REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE 30-Sep 97	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Lip Tracking for Audio-Visual Speech Recognition

5. FUNDING NUMBERS

6. AUTHOR(S)
Robert August Kaucic Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Oxford

8. PERFORMING ORGANIZATION REPORT NUMBER

97-031

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
THE DEPARTMENT OF THE AIR FORCE
AFIT/CIA
2950 P STREET
WPAFB OH 45433

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

19971006 142

14. SUBJECT TERMS

DTIC QUALITY INSPECTED 4

15. NUMBER OF PAGES
160

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

Robert August Kaucic Jr.

Merton College

Doctor of Philosophy

Trinity Term 1997

Lip Tracking for Audio-Visual Speech Recognition

Abstract

Human speech is conveyed through both acoustic and visual channels and is therefore inherently multi-modal. Further, the two channels are largely complementary in that the acoustic signal typically contains information about the manner of articulation while the visual signal embodies knowledge of the place of articulation. This orthogonal nature of the audio and visual components has enticed researchers to develop audio-visual speech recognition systems that have been shown to be robust to acoustic noise. A fundamental requirement of automatic audio-visual speech recognition is the need for real-time tracking; however, this necessity has been largely ignored by the lipreading[1] community. This work presents a new approach for tracking unadorned lips in real time (50 fields/sec). The tracking framework presented combines comprehensive shape and motion models learnt from continuous speech sequences with focused image feature detection methods. Statistical models of the grey-level appearance of the mouth are shown to enable identification of the lip boundary in poorly contrasted grey-level images. The combined armoury of the these modelling approaches permits robust, real-time tracking of unadorned lips.

Isolated-word recognition experiments using dynamic time warping and Hidden Markov Model-based recognisers demonstrate that real-time, contour-based, lip tracking can be used to provide robust recognition of degraded speech. In noisy acoustic conditions, the performance of recognisers incorporating visual shape parameters are superior to the acoustic-only solutions, providing for error rate reductions up to 44%. Further experiments using individual shape components suggest that the recognition information in the outer lip contour is concentrated in three shape parameters, approximately corresponding to 'ah', 'ee', and 'oh'.

In order to capture the linguistic information carried by the teeth and tongue, more capable trackers are also introduced which exploit information-rich colour images. Feature detectors, which employ Bayesian discriminant analysis techniques on colour images, provide for fast, accurate, identification of the boundary between the lips and their surround. The result is a robust tracker capable of tracking both the inner and outer lip contours. This tracker permits more detailed measurements to be made about the teeth and tongue, and serves as a foundation for further exploration of the benefits of lipreading.

---

[1]The term "speechreading" more accurately describes the process of using visual information to understand speech; however, "lipreading" is the more commonly accepted term.

Lip Tracking for Audio-Visual Speech Recognition

Robert August Kaucic Jr.
Robotics Research Group
Department of Engineering Science
University of Oxford
1997

This thesis is submitted to the Department of Engineering Science, University
of Oxford, for the degree of Doctor of Philosophy. This thesis is entirely my
own work, and except where otherwise indicated, describes my own research.

Robert August Kaucic Jr.
Merton College

Lip Tracking for Audio-Visual Speech Recognition

Abstract

Human speech is conveyed through both acoustic and visual channels and is therefore inherently multi-modal. Further, the two channels are largely complementary in that the acoustic signal typically contains information about the manner of articulation while the visual signal embodies knowledge of the place of articulation. This orthogonal nature of the audio and visual components has enticed researchers to develop audio-visual speech recognition systems that have been shown to be robust to acoustic noise. A fundamental requirement of automatic audio-visual speech recognition is the need for real-time tracking; however, this necessity has been largely ignored by the lipreading[1] community. This work presents a new approach for tracking unadorned lips in real time (50 fields/sec). The tracking framework presented combines comprehensive shape and motion models learnt from continuous speech sequences with focused image feature detection methods. Statistical models of the grey-level appearance of the mouth are shown to enable identification of the lip boundary in poorly contrasted grey-level images. The combined armoury of the these modelling approaches permits robust, real-time tracking of unadorned lips.

Isolated-word recognition experiments using dynamic time warping and Hidden Markov Model-based recognisers demonstrate that real-time, contour-based, lip tracking can be used to provide robust recognition of degraded speech. In noisy acoustic conditions, the performance of recognisers incorporating visual shape parameters are superior to the acoustic-only solutions, providing for error rate reductions up to 44%. Further experiments using individual shape components suggest that the recognition information in the outer lip contour is concentrated in three shape parameters, approximately corresponding to 'ah', 'ee', and 'oh'.

In order to capture the linguistic information carried by the teeth and tongue, more capable trackers are also introduced which exploit information-rich colour images. Feature detectors, which employ Bayesian discriminant analysis techniques on colour images, provide for fast, accurate, identification of the boundary between the lips and their surround. The result is a robust tracker capable of tracking both the inner and outer lip contours. This tracker permits more detailed measurements to be made about the teeth and tongue, and serves as a foundation for further exploration of the benefits of lipreading.

---

[1]The term "speechreading" more accurately describes the process of using visual information to understand speech; however, "lipreading" is the more commonly accepted term.

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Andrew Blake, for his invaluable direction and assistance in the development of this work. Thank you also to many of the gifted members of the Robotics Research Group — Benedicte Bascle, Barney Dalton, Colin Davidson, Michael Isard, David Lee, John MacCormick, Ben North, David Reynard, Simon Rowe, Josephine Sullivan, and Andrew Wildenberg — for their assistance, fruitful discussions, and for writing much essential software.

As an officer in the U.S. Air Force, I am also deeply indebted to the U.S. Air Force and the American taxpayers whose financial support made this work possible.

I would also like to extend my love and apologies to my two children, Catherine and Robert. To Catherine, you may never know how difficult it was for me to return to work each night as you acceptingly said, "Daddy work." And to Robert, I hope that one day you will be able to laugh as you recount how you spent the first nine months of your life sleeping in the dining room.

Finally, I would like to thank my loving wife, Susan, for everything — her willingness to learn to lipread, her caring for me, her raising of our children, and most importantly, her undying love and support.

Contents

1

Introduction

Since verbal communication is the principal method of conveying information between humans, the possibility of communicating with computers through simple verbal interaction presents an opportunity to profoundly change the way humans interact with machines. Voice interactive systems will relieve users of the burden of entering commands via computer keyboards and mice, and could prove indispensable in situations where the operator's hands are occupied, such as when driving a car or operating machinery. Much research has focused on the development of spoken language systems, and rapid advances in the field of automatic speech recognition have been made in recent years [32, 96, 147]. Although progress has been impressive, researchers have yet to overcome the inherent limitations of purely acoustic-based systems, particularly their susceptibility to environmental noise. Such systems readily degrade when exposed to time-varying or unpredictable noise as might be encountered in a typical office environment with ringing telephones, background radio music, and disruptive conversations. Their performance also drops in more benign situations, such as inside moving automobiles or when the signal is transmitted across telephone lines. In tests conducted by the Advanced Research Projects Agency (ARPA), error rates on state-of-the-art acoustic recognisers more than doubled when presented with speech distorted by telephone channels [96].

If speech recognition systems can be made to function effectively in noisy environments, then voice-interactive technology can be extended to a wide range of application domains. For instance, an application of current commercial interest is the ability to operate and

control a car-phone by voice, enabling, for example, hands-free dialing. Similarly, speaker identification/verification techniques can be used in security applications, such as access control. Eventually, security systems may even combine visual face recognition with the voice pattern matching. Interfaces to standard office equipment such as computers, photocopiers, and fax machines could also be improved by the employment of user-friendly, voice-interactive front ends. The principal obstacle to the utilisation of current speech recognisers in these environments is their poor performance in the presence of interfering noise.

To enable operation in adverse environments, acoustic solutions typically use noise compensation methods during pre-processing or recognition. The pre-processing approaches often use noise masking, noise cancellation, spectral subtraction, or adaptive filtering techniques to remove the additive noise power from the signal [80, 21, 103]. Hidden Markov Model (HMM) decomposition, where separate models are used for the clean speech and noise, is a common method used to provide compensation during recognition [135, 51, 52]. While these approaches have proven to be effective, they ignore a basic truth, that is, the multi-modal nature of human communication. This research attempts to exploit this reality by using visual information, in the form of parameters describing the shape of the lips, to improve upon acoustic speech recognition performance. Further, although only limited attention is given to noise compensation methods in this work, it should be noted that the employment of noise compensation techniques and the inclusion of visual information are not mutually exclusive. Rather, the beauty of the visual signal is that it can be utilised in conjunction with any of these acoustic compensation strategies.

It is well known that human speech perception is enhanced by seeing the speaker's face and lips — even in normal-hearing adults [43, 116]. Sumby and Pollack [129] have shown that visual information enhances speech understanding, especially in noisy environments. Further, several researchers [108, 130] have demonstrated that the primary visible articulators (teeth, tongue, and lips) provide useful information with regard to the place of articulation and Summerfield [130, 131] concluded that such information conveyed knowledge of the mid- to high-frequency part of the speech spectrum — a range readily masked by noise.

Motivated by this complementary nature of the visual information, researchers have recently developed audio-visual speech recognisers which have proven to be robust to acoustic noise [105, 128, 16, 26, 1]. Although the field of audio-visual speech recognition shows great promise, it is still in its infancy, even in comparison with acoustic speech recognition which after fifty years of research has only recently resulted in commercially available systems

like DragonDictate from Dragon Systems and VoiceType Dictation from IBM. There are many hurdles that must be overcome before audio-visual speech recognition systems become commercially viable.

If audio-visual recognition systems are to be effective, they must be capable of tracking the lips (inner or outer contour, or both) and be robust to head movements and variations in lighting and pose. The tracker must be able to follow the lips for an extended period of time, at a minimum the length of one interactive session, which in the case of car-phone applications may be several hours. In addition, real-time tracking is essential as applications dealing with man-machine interfaces do not afford an opportunity for off-line processing.

The need for accurate tracking of the lips, however, is only one step in the larger problem of extracting linguistically relevant information from the visual channel. In order to provide accurate recognition, the tracking must yield information that can be used to discriminate between the various recognition units (words, phonemes, tri-phones). Although it is known that humans supplement their understanding of speech using visual cues — which include visual movements contributing directly to production of speech, such as the positioning of the lips, teeth, and tongue, as well as more peripheral movements like head nods and eyebrow movements [99, 94, 9, 23] — it is not known which visual recognition features[1] are the most beneficial from a machine recognition perspective.

Lastly, a problem that has proved to be far more challenging than may have been anticipated is the intelligent integration of the audio and visual channels. On the surface, integration of the two modalities appears to be a straightforward data fusion problem. However, complications arise because the relative importance of the channels varies as a function of the spoken word. For example, "me" and "knee" are more easily distinguished visually than aurally, while discriminating between "me" and "pea" is principally an acoustic charge. Further, the integrity of the acoustic channel is strongly influenced by the presence of interfering noise. Accordingly, audio-visual integration strategies should be able to adapt to changing noise conditions and appropriately weight the two information sources according to their linguistic relevance.

The many challenges associated with the audio-visual speech recognition problem pro-

---

[1]It is unfortunate that the word "feature" is used to describe markedly different entities by the speech recognition and vision communities. In recognition parlance, a feature is a representation of a signal which compactly captures its information content (typically spectral representations in the case of acoustic signals and often via geometrical measures like height, width, area, or shape parameterisations, for the visual signals). In the vision community, a feature is a distinguishable point in an image, such as an edge, valley, corner, or even a boundary between two regions, like the lips and skin. When the intended meaning is ambiguous, *recognition features* will be used to indicate the compact representation of information, while *image features* will designate distinguishing points in images.

vide a rich source of interesting areas for research. This thesis addresses several of these problem areas with the primary emphasis being on providing solutions to the difficult lip-tracking problem.

The problem of accurately tracking rapidly moving, articulating lips is a formidable one. The task is compounded by the real-time constraints imposed by the target application. Currently, other than the work presented here, Petajan et al. [106, 107] possess the only tracker capable of tracking unadorned lips in real time. Their tracker, as well as the work of many others, is detailed in chapter 2. In contrast with other tracking work, the lip trackers described in this thesis use a dynamic contour tracking framework [12, 15] to attain real-time performance. An overview of these lip trackers and the dynamic contour framework is given in chapter 3. In chapter 4, it is shown that the dynamic contour tracking framework is well suited for tracking rapidly moving, articulating lips from profile views. It is further demonstrated that highly accurate frontal lip tracking is attainable if lipstick is used to enhance the contrast of the lips. However, when applied to unadorned lips, the computation-efficient, edge-based feature detectors used in the trackers are shown to be ill-suited for tracking the weakly contrasted lips. Statistical models of the grey-level appearance around the lips are employed which capture the salient information for identifying the lip boundary. When incorporated into the tracking framework, the result is accurate tracking of unadorned lips. Although the use of statistically-based feature detectors is itself not new [34, 122], this is the first reported use of their employment in real-time tracking problems.

In order to evaluate the extent to which lip contour information can be used to aid speech recognition, two audio-visual speech recognisers are developed. The two recognisers, one which uses a dynamic time warping (DTW) pattern matching algorithm and the other which uses continuous density Hidden Markov Models (HMMs), are explained in chapter 5. Visual *shape* parameters are obtained from the tracked lip contour by projecting the lip outline onto recognition bases. Several recognition experiments which were conducted on isolated-word vocabularies with and without added Gaussian acoustic noise are presented in chapter 6. These experiments demonstrate that shape parameters obtained from accurately tracked lip contours can be used to provide robust speech recognition in the presence of high levels of interfering noise. Further experiments using individual shape components suggest that the recognition information in the outer lip contour is concentrated in three shape parameters, approximately corresponding to 'ah', 'ee', and 'oh'. When the acoustic channel is degraded, the visual information significantly enhances recognition performance — reducing error rate up to 44%. Improvements occur even when the acoustic recogniser is trained and tested at the known noise level. However, in clean acoustic conditions, the

shape parameters provide only a slight reduction in error rates ($<$ 8%). Additional visual information in the form of knowledge of the inner mouth region, to include the teeth and tongue, may be needed to increase performance in these environments.

Towards this end, more capable trackers are developed in chapter 7 which make use of the increased discriminating potential of colour images of the face. First, a novel application of Fisher's Linear Discriminant Analysis [46] is presented which enables accurate identification of the lip-skin boundary and is shown to be robust to environmental variations. Further, since the learning of the Fisher discriminant is done off-line, outer lip contour tracking can still be accomplished in real time on general-purpose workstations (Silicon Graphics Indy R4400 200 MHz). Next, accurate demarcation of the inner mouth contour is attained despite considerable variations in the appearance of the mouth due to movements of the teeth and tongue. Mixtures of multi-variate Gaussians enable precise modelling of the colour intensities inside the mouth. The resultant inner-outer lip contour tracker permits extraction of information from the image data inside the mouth, thus enabling more detailed judgements to be made about the presence and position of the teeth and tongue.

2

Background

2.1 Human Lipreading and Speech Perception

In developing systems that attempt to mimic human capabilities, such as hearing and seeing, it is often helpful to study those who do it best — humans. It is not surprising that deaf and hearing-impaired individuals use lipreading as their primary source of information for speech communication [42]. Nor is it surprising that visual cues improve speech perception in acoustically noisy environments [129]. What may be surprising is the extent to which seeing a speaker's face and lips affects speech perception for normal-hearing people in clean acoustic environments. It is widely accepted that sound is the primary instrument for human speech recognition (and one need only turn down the volume on the nightly newscast to verify this). Despite the temptation to relegate lipreading to a "back-up" system when the audio system fails or is degraded, researchers have shown that visual cues enhance speech perception even in clean acoustic environments [130, 93, 116]. Reisberg et al. [116], convinced that lipreading was more than a back-up system, devised tests where the audio signal was easy to hear, but hard to understand. They exposed subjects with normal hearing and no lipreading training to foreign languages, a speaker with a foreign accent, and a semantically complex message. In all experiments, despite noise-free audio signals, those who saw the speaker's face recognised a greater percentage of the words [116].

In a classic study, McGurk and MacDonald [95] demonstrated that when presented with conflicting aural and visual stimuli, listeners often reported *hearing* neither the aural

or visual stimulus, but a blend of the two. In their experiments, subjects were shown a video with a speaker mouthing "ga" which had been dubbed with audio corresponding to "ba". When asked to report what they heard, subjects reported hearing "da". This audio-visual blend, now commonly referred to as the "McGurk effect", indicates that visual information affects speech interpretation even when the acoustic signal is clear and unambiguous.

Given that the visual cues can complement acoustic ones, it is important to determine what information they provide and assess situations where they can be of most benefit. Sumby and Pollack [129] found that the contribution of visual information to speech intelligibility increased as the signal-to-noise ratio (SNR) of the audio signal decreased, primarily due to the poor intelligibility of speech at low SNRs. Campbell [27], extrapolating from the work of Sumby and Pollack, deduced that seeing the talker's face was equivalent to a 15 dB increase in the acoustic signal-to-noise ratio. Summerfield et al. [130, 131] provide an explanation. They assert that movements of the visible articulators (teeth, tongue, and lips) convey information on the *place* of articulation. This indicates spectral detail in the mid- to high-frequency region of the speech spectrum — a region readily masked by noise. Conversely, the acoustic signal, which denotes movements of the hidden articulators (larynx and velum), conveys the *manner* (voicing and nasality) of the speech. These movements are typically associated with the intense low-frequency part of the speech spectrum and are less susceptible to noise. Thus, the audio and visual signals are complementary, with the visual signal being most beneficial in the region where the acoustic signal is most vulnerable to the deleterious effects of the noise.

In related experiments, Brooke, McGrath, and Summerfield [25, 94, 131] investigated the comparative recognition rates of individuals presented with differing amounts of visual stimuli. Subjects were shown images of the speaker's entire face, the lips and teeth only, and the lips only, and asked to identify the vowel present in a /b/-Vowel-/b/ context. As was expected, the best performance (78% recognition rate) was achieved when the observers saw the entire face. Recognition rates dropped to 56% using the lips and teeth and 50% for the lips only. Their results suggested that, although visual information from the lips alone contains reliable recognition information, it may be necessary to incorporate additional cues, such as the teeth and tongue, to approach the recognition potential of the entire face. In similar studies using French nonsense words, Benoit et al. [9, 84] confirmed the above findings, further concluding that the "lips alone carry on average two-thirds of the speech intelligibility carried out by the whole natural face."

The findings of these psychological studies highlight the multi-modal nature of human communication and demonstrate the importance of lipreading in speech perception. They

also serve as the impetus for the incorporation of lipreading information into automatic speech recognisers. As Summerfield says, "The potential improvements to be gained from automatic lipreading are sufficiently large for speech technologists to be exploring image-processing algorithms to bolster the performance of acoustical recognisers." [131]

2.2 Computer Lipreading Systems

Although it has been known for sometime that human speech perception is a multi-modal process, it is only recently that researchers have begun to explore the potential benefit of incorporating visual information into acoustic speech recognisers. The first serious work in the area was performed by Petajan and Brooke [104, 24, 105] (and later [106, 26, 133]) followed shortly by Finn and Montgomery [47], but recently many others have entered the field, including Bregler et al. [16, 17, 18, 19], Stork et al. [128], and Benoit et al. [9, 1]. As in acoustic recognition systems, there are many different ways in which to classify the various lipreading systems. Such systems can be classified according to the method used to track the lips, the visual recognition features extracted, the recognition method (DTW, HMM, NN, etc.) used, or the audio-visual integration strategy employed. Furthermore, they may be speaker-dependent or speaker-independent (both with regard to tracking as well as recognition) and may operate on isolated-word or continuous speech databases. Despite the large variability among the many systems, there are several fundamental issues that each system must address. Foremost among them are the method used to track the lips and the type of visual information extracted.

If audio-visual recognition systems are to be effective, they must be capable of tracking the lips (inner or outer contour, or both) and reasoning about the presence/absence and position of the teeth and tongue. Ultimately, they should adequately handle unconstrained speakers who may be moving around freely, nodding or rotating their heads. They should also be robust to variations in lighting and shadowing. Furthermore, in order to provide accurate recognition, the tracking must yield information that can be used to discriminate among the various recognition units (words, phonemes, tri-phones). This extracted visual information must also be intelligently integrated with the acoustic features, presumably in proportion to the information content of each channel. Finally, all of this should, of course, be accomplished in real-time or near real-time in order that practical use may be made of such systems.

Although there is currently no consensus in the speechreading community for what the best strategy is for tackling these issues, it is instructive to look at the methods used by

various researchers in addressing these fundamental problems. Since there are no standard or commercial audio-visual databases against which the various systems can be measured, it is difficult to make meaningful comparisons between them. However, it is nonetheless useful to take a detailed look at some of the more representative systems, noting their strengths and weaknesses.

2.3 Lip Tracking and Visual Feature Extraction

The problem of accurately tracking rapidly moving, articulating lips is a formidable one. The task becomes even more difficult if the computational time constraints required by the intended applications — audio-visual speech recognition, lip synchronisation for animation, expression recognition, etc. — are considered. Currently, other than the work presented here, Petajan et al. [106, 107] possess the only tracker capable of tracking unadorned lips in real time on general-purpose hardware. The need for the tracking of the lips, however, is only one step in the larger problem of extracting linguistically relevant information from the visual channel. In essence, feature extraction is the process of reducing the high-dimensional image data down to compact sets of *features* that adequately represent the information content of the visual signal. Choosing the most suitable representation of the visual information remains an open research question. In general, there are three main feature extraction approaches, those that use the pixel intensity information, those that use image flow or motion, and those that use lip contour information. Naturally, the feature extraction method employed greatly influences how lip tracking is accomplished.

2.3.1 Pixel-based systems

Owing to the difficulty of accurately tracking the lips, and the belief that it is the recognition engine that should determine the informative visual features, many researchers [142, 26, 101, 124] extract their features directly, or indirectly, from the grey-level pixel data. Vector quantisation [125] or principal components analysis [26] are typically used to reduce the high-dimensional image data down to a more manageable size. A major strength of this approach is that most of the information about the primary visible articulators — the teeth, tongue, and lips — is retained. The downside to this approach is that these systems tend to be highly susceptible to changes in lighting, viewing angle, and speaker head movements. Such systems may be optimal in certain settings where there is limited head movement relative to the camera and where the lighting can be carefully controlled. Indeed, impressive audio-visual recognition results have been achieved in controlled lighting conditions by several

researchers [17, 133] using the pixel intensities directly. Such an approach may still prove effective in more natural settings, although recognition performance will likely depend on the ability of the recognition engine to generalise over lighting changes and compensate for tracking errors.

Petajan and Brooke [104, 105] were the first to demonstrate that visual information from lipreading could be used to improve the performance of acoustic speech recognisers. They conducted speaker-dependent, isolated-word experiments on four separate speakers. Their system consisted of a commercial acoustic recogniser and an image processing system. Images of speakers' faces were thresholded with a manually set value to produce binary images. The threshold was chosen so that no dark mouth regions were present when the mouth was shut. They measured the distance between each speaker's nostrils and their mouth. The location of the speaker's nostrils was identified using region matching against a stored nostril template. They assumed that the distance between the mouth and the nostrils remained relatively constant and thus the region around the mouth was located using this known distance. Vector quantisation (VQ) was used to represent these binary mouth images as a 256-symbol codebook.

Recognition was performed in two separate stages. Features (mouth images) were extracted from unknown vision sequences as described above. They were then represented as a series of codebook symbols and matched against two representative samples of each word. The unknown word was then determined using dynamic time warping with a Euclidean distance measure. For the combined audio-visual system, the commercial acoustic recogniser identified the top two most likely candidates for the unknown and then the visual recogniser decided between the two.

Separate tests were conducted on two "clean" (no artificially added noise) databases — the 10 digits and the 26 English letters. Recognition error rates on the alphabet ranged from 28–45% for acoustic-only, 20–28% on vision-only, and 11–29% for the combined audio-visual system. Their key finding was that the visual information did indeed contain important recognition information that could supplement acoustic-only systems. Results for both databases and for all four speakers confirmed this.

Similarly, Silsbee et al. [126, 124] used VQ codebooks to represent the pixel data; however, rather than using binary images of the face, they used smoothed grey-levels. Visual features were acquired by extracting an 80×80 pixel array from a predefined region of interest in the image which bounded the speaker's mouth. Histogram flattening and left-to-right balancing were used to reduce some of the effects of the variations in lighting. VQ was used to represent these 6400-dimensional pixel intensity vectors as a 32-symbol code-

book. Codebook entries were assigned to the test images by choosing the codebook image with the minimum squared pixel-intensity difference between it and the test image. They conducted isolated-word recognition experiments on two different databases. In one series of experiments [126], they used 22 English consonants in /a/-Consonant-/a/ context and in another [125], a 500-word database. Their audio-visual recognition system was driven by the fact that their hardware could not support real-time acquisition of both audio and visual data. As a result, the data were acquired on separate machines without any means of synchronising the audio and visual sequences. Twelve audio features were extracted using Hermansky's Perceptual Linear Prediction (PLP) [65]. As with the visual data, these too were represented as three 32-symbol codebooks. Actual recognition was accomplished using discrete-density, left-to-right HMMs. The audio and visual sub-systems were phoneme based, but completely independent (because of the synchronisation problem) which permitted different state paths through the HMMs for the audio and visual sequences. In order to combine the two sub-systems, the class-conditional probabilities were weighted by a user-defined constant λ, that is, $\Pr(\text{Word}|A, V) = \Pr(\text{Word}|A)^{\lambda} \times \Pr(\text{Word}|V)^{(1-\lambda)}$. Average error rates on the 22 consonants using clean speech were 4% for audio only, 52% visual only, and 4% for the audio-visual system. Average error rates on the 500-word database using clean speech were 40% for audio only, 78% visual only, and 22% for the audio-visual system. Using noisy speech (10 dB SNR), error rates were 50% for audio only, 78% for visual only, and 40% for the audio-visual system.

2.3.2 Flow-based tracking

A second method of extracting features from images of speakers is based on the premise that humans are more sensitive to *motion* than to static scenery. It is therefore believed [92] that the motion of the lips contains more useful recognition information than the physical outline or the grey-level intensities of the mouth region. In particular, lip velocity is believed to aid in locating word and/or syllable boundaries. A limitation of this approach is that computationally expensive procedures like optical flow analysis and morphological operations are used to extract the lip velocities which restricts their use to applications that do not require real-time performance. Further, typically only coarse measurements of mouth motion are obtained, whereas more precise estimation of the motion is desired.

Mase and Pentland [92] used Horn and Schunk's gradient method [69] to estimate the optical flow around the mouth. Specifically, they computed the horizontal and vertical velocity of the mouth in four rectangular regions around the mouth. They obtained two features from these velocities — a measure of mouth opening/closing movements and a

measure of the elongating/contracting of the sides of the mouth.

Visual only, multi-speaker (3-speaker) recognition experiments were performed using connected words (digits). Recognition was accomplished using a weighted Euclidean distance measure between sequences of the two features. All of the utterances (digit sequences) were linearly time-warped to a standard length of sixteen samples. The "distance" between two sequences was computed by weighting the squared difference between each feature in the sequence by the ratio of the eigenvalues of the two features (determined through principal component analysis) and summing over the entire sequence. Average error rate for the three speakers was 24%, although the test set consisted of only four word sequences (21 total words).

No special equipment was used to fix the position of the speakers' heads, although speakers were asked to rest their heads against a wall to limit extraneous movements. The authors recognised that global head movements not related to speech production could contaminate their data, although they felt that this deficiency could be overcome by compensating the local optical flow with information about global movements.

Mak and Allen [89] investigated the use of lip motion to improve segmentation of noisy speech along syllabic boundaries. They noted that in images of faces, the gap between the upper and lower lips (inner mouth region) is often the darkest part of the image. Through the use of morphological erosion, image subtraction, binary thresholding, and cluster analysis, they were able to locate the centre of the lips. They extracted velocity measures similar to the opening/closing and elongation/de-elongation used by Mase and Pentland [92] using an exhaustive block matching algorithm. For each vision frame, a single resultant velocity, V_r, was computed using the two velocity measures, V_x and V_y, by $V_r = \sqrt{V_x{}^2 + V_y{}^2}$. The peaks and troughs of $V_r(t)$ were then used to identify syllable boundaries in continuous speech. The authors claim that using a combination of $V_r(t)$ and the acoustic signal resulted in a 10% reduction in segmentation errors; however, a notable limitation of this system is that it uses only local image motion and hence global head movement can be erroneously detected as lip motion.

Despite recent findings [55, 17, 57] that motion does indeed carry important linguistic information, flow-based tracking has fallen out of favour in the speech reading community. Rather than using flow-based methods for obtaining the lip velocity, tracking is used to determine the time-varying lip position, and delta positions are used as velocity [17, 55].

2.3.3 Dot tracking

Since human lipreaders rely heavily on the positioning of the lips [25, 131], extraction of *shape* parameters from the lip outline presents a third alternative to feature extraction. In the simplest case, reflective dots on the face can be tracked over time to obtain geometric measures like height, width, area, and circumference of the mouth opening. However, more advanced model-based tracking can be used to acquire lip deformations corresponding to particular actions, or sounds, such as the lip rounding of 'oo' or the curling of the lip corners in 'ee' (section 6.1). One obvious limitation of this approach is that information on the positioning of the tongue and teeth is lost unless additional steps are taken to retain it. Furthermore, since there is not a prominent edge at the boundary between the lips and face [145, 100], accurate tracking of the lip outline is itself a formidable problem. To overcome this, early researchers [47, 22, 128] placed reflective dots around the speakers' mouths in order to obtain accurate measurements of lip shape parameters. The principal advantage of using shape parameters as inputs to the recognition engine is that they are inherently invariant to changes in illumination and can be made robust to head movements.

Finn and Montgomery [47] were the first to investigate the use of shape features in automatic speech reading. They conducted speaker-dependent, isolated-word recognition experiments on 23 English consonants in /a/-Consonant-/a/ environment. Tracking of the mouth outline was accomplished by recording the positions of twelve highly reflective dots placed around the speakers' mouths. Fourteen distance measurements were computed using the recorded positions. Each consonant in the database was said twice, once for use as a reference template and once for testing. All utterances were truncated by hand to 29 frames (with duration 29/30 sec) and aligned such that the temporal centre was at the maximum vertical opening of the mouth.

Recognition was accomplished using direct (no time warping) calculation of a weighted Euclidean distance between the test utterance and the 23 reference templates. The token yielding the smallest total distance was identified as the recognised token. With equal weighting for all 14 distance measurements, recognition error rates of 60% were achieved using only the visual data . Allowing for the fact that it may be impossible to distinguish consonants in the same viseme group (ie. 'b' and 'p', 'd' and 't', etc.) using vision alone, and thus counting "apa" correct for both "aba" and "apa", the error rate dropped to 22%. When experimentally-determined "optimal" weighting was used in the Euclidean distance measure, the error rates dropped to 13% on the consonants and 5% on the viseme groups.

The authors mentioned the use of a commercial acoustic speech recogniser on utterances

at eight different SNR ratios and discussed ways of incorporating their visual features into the acoustic recogniser, but no results of either were presented.

2.3.4 Model-based tracking

The success of Finn and Montgomery's work using geometric features and recent advances in model-based tracking have led many researchers to explore model-based techniques for lip tracking. Most "lip trackers" build upon the pioneering "snake" approach of Kass and Terzopoulos [75] or the "deformable template" methods of Yuille et al. [146, 144]. The essence of the model-based tracking approach is the incorporation of prior *shape* knowledge about the object to be tracked. Identification of the outline of an object is formulated as an energy minimisation problem: internal energy terms are used to impose continuity and smoothness constraints on the deforming contour, while external energy terms serve to guide the contour to salient features in the image, such as edges and valleys. Kass' snake approach provides only general, or soft, constraints on the allowable shapes, encouraging the contour to be elastic, or biasing it towards long, thin shapes, while Yuille's deformable template approach imposes hard constraints on permissible shapes by explicitly parameterising the contour.

The audio-visual speech recognition systems developed by Bregler and his colleagues [16, 17, 20], which are some of the most comprehensive systems developed to date, use Kass's snake approach with shape constraints imposed on possible contour deformations. They track only the outer lip contour while restricting the allowable lip shapes to lie on a manifold which is learnt from training sequences of lip shapes. Early versions employed linear shape spaces [16]; however, a more recent tracker [18, 19] permits the lip contour to lie along a non-linear manifold. The latter approach also enables non-linear interpolation between successive images of the mouth to permit synchronisation of information from the audio and visual channels, which run at different rates. Within this framework, the lip outline is represented by 40 evenly spaced points along the contour; however, rather than using the tracked lip position as input to their recognition engine, it serves only to provide an image location about which a 24×16 matrix of pixel intensities is extracted. Visual processing consists of reducing dimensionality of the pixel data from 384 (24×16) to 10, using principal component analysis. Acoustic parameters consist of 8 cepstral coefficients plus a ninth feature corresponding to the average acoustic energy in each frame.

They investigated multi-speaker (6-speaker) audio-visual recognition of connected words (German letters). They used a multi-state Time Delay Neural Network recogniser which was trained using the concatenation of the 9 audio features, the 10 principal components

from the grey-level intensities, and an additional 10 "delta" features from the change in grey-level intensities. They tested their system using clean audio data, audio data with additive noise recorded inside a moving car, and additive crosstalk data simulating the "cocktail party" effect. On the clean data their acoustic-only system yielded an error rate of 11% and their combined audio-visual system 10%. For the audio data with additive car noise at a SNR of 20 dB, the error rates were 56% and 48%. There was a similar improvement for the 15 dB SNR crosstalk-corrupted audio, where the combined system resulted in an improvement from an error rate of 67% to 46%. In addition, they also attempted recognition using features extracted from the lip contour. They provided no specific numbers, but concluded that the outline of the lip alone was not distinctive enough to give reliable recognition performance. This is most likely due to their use of image forces consisting of only grey-level gradients, which are known to be inadequate for identifying the outer lip contour [144, 86, 76].

One aim of this thesis was to test the claim that the lip contour was "not distinctive enough to give reasonable recognition performance" [19]. It is our belief that accurately tracked lip contours are a rich source of information for audio-visual speech recognition — a belief supported by recognition experiments presented in chapter 6. Moreover, we demonstrate that tracking can be accomplished at real-time rates (50 Hz) — addressing a compelling requirement of audio-visual speech recognition.

Several researches, Stork et al. [63, 62, 64], Rao et al. [115], and Silsbee et al. [29], employ pared-down versions of Yuille's iterative gradient descent minimisation deformable templates to track both the inner and outer lip contours. In the interest of computational efficiency, rather than using the full complement of image potential fields and heuristic constraints (penalty terms) as proposed by Yuille [144], an abbreviated set is typically used [63, 115]. The result of foregoing special energy terms to account for the appearance and disappearance of the teeth and to compensate for the lack of identifying edges along the lower lip is a reduction in processing time from 5 minutes per frame [144] to around 1 second per frame. However, the computational savings come at the cost of decreased tracking accuracy.

Recently, Luettin et al. [86, 87] have achieved some success using the Point Distribution Models (PDMs), also referred to as Active Shape Models (ASMs), of Cootes and Taylor [33, 83] to identify the inner and outer lip contours. These models allow objects to be represented as a connected series of image points — polygons. The principal modes of shape variation of objects are learnt from hand-labelled training images. Deformations of the object model (template) are restricted to lie in a shape space derived from principal components analysis

on the training data. An iterative refinement (gradient descent) algorithm is used to deform the template to best fit the feature support found in the image.

In order to overcome the lack of consistent identifying edges along the lip contours, Luettin et al. [86, 87] use models of the grey-level intensities along the inner and outer lip contours to identify image features corresponding to the lip boundary. However, the intermittent presence of the teeth results in template profiles for the inner mouth contour with large variances that, at times, are not sufficient to accurately pinpoint the lip contour, resulting in tracking errors.

2.3.5 Real-time tracking

Despite the success achieved using model-based tracking, it is surprising that researchers have failed to consider the real-time tracking requirement of the audio-visual speech recognition application. It is our view that the real-time constraint is more than just a call for efficient algorithms. Indeed, computational power doubles approximately every eighteen months permitting more and more systems to fall under the real-time umbrella in only a few years. However, in some cases the computational requirements of the lip trackers employed are hundreds or even thousands of times slower than real time, meaning that it will probably be at least a decade before such techniques can be employed in practical systems. Furthermore, there is still a great deal of ongoing research as to which features most efficiently capture the relevant linguistic information, and it is likely that as more is learned about this, additional processing will be necessary to capture the informative parts of the visual signal, making it even more unlikely that systems that completely ignore time considerations will ever be realised in practice. Our approach has been to start with a real-time framework, and then add additional, focused processing where needed. With this approach, even when the processing requirements exceed the current hardware capabilities, resulting in slower than real-time performance, real-time performance can be regained in short order with the inevitable hardware advances. It is our belief that developing research platforms using this design approach enforces consideration of issues that are easily otherwise overlooked.

A second, possibly more compelling, reason for operating within real-time constraints is that it imposes a more rigorous standard of acceptability. For example, one group of researchers [86] accessed the performance of their lip tracker by classifying the contour fit to the lips as good, adequate, or a miss. Although the 6% miss rate that they reported might be considered quite successful, had the tracker been evaluated on even a 30-second sequence of continuous speech (1500 fields), it is likely that tracking performance would

be re-assessed as unacceptable. The ability to track in real time permits evaluation of tracking performance over an extended period of time, which frequently reveals problems not observed in short tracked sequences. In contrast, trackers which are unable to operate in real time, typically must run off images stored on disk. In such cases, the shear magnitude of disk space required (over 60 MB per 10 second sequence) often precludes evaluation on anything other than very short sequences. In this thesis, all of the dynamic contour trackers presented have been tested on "live" speech, principally, hours of video recordings.

Currently, other than the work presented in this thesis, the only tracker capable of tracking unadorned lips in real time is the inner contour tracker of Petajan et al. [106, 107, 56]. Instead of relying on prior models of deforming lips, they cleverly utilise the fact that human nostrils represent two dark spots on the face. If the nostrils can be seen, anatomical constraints can be used to concentrate image processing on the eye region (to determine head tilt angle) and on a rectangular window around the mouth. Colour thresholds are used to identify the black area in the inner mouth region and neighbouring pixels are compared against "teeth coloured" templates. A contour is then grown around the area identified as the inner mouth. Tracking has been shown to be robust to head tilt and speaker facial hair. The only drawback of this system is that it relies on having a clear view of the nostrils, which is available in applications where the camera can be mounted to look up at the speaker, but which may not be satisfied in general viewing conditions of the face, or even typical fronto-parallel views.

2.4 Recognition and Integration Strategies

The recognition algorithms used for audio-visual speech recognition are essentially the same as the pattern matching approaches — dynamic time warping, neural networks, and Hidden Markov Models — used for acoustic speech recognition. A problem that has proved to be far more challenging than may have been anticipated is the intelligent integration of the acoustic and visual channels. Ideally, the information from the two channels should be integrated in proportion to their information content. For instance, when the acoustic channel is degraded, one would expect to rely more heavily on the visual channel. However, even in clean acoustic conditions, words that are better distinguished by place of articulation, such as "mow" and "no", should naturally make optimal use of the visual channel, while those better distinguished by manner of articulation, such as "me" and "pea", might rely more heavily on the acoustic channel. The challenge of audio-visual integration is how best to combine the two channels making optimum use of the informative aspects of each

channel. The problem is further complicated when one expects human-like performance of the audio-visual system; ideally the recogniser should seamlessly adapt to variations in the level and type of interfering noise, and provide audio-visual recognition rates that exceed those obtained using audio-only or visual-only data over a broad range of noise conditions.

2.4.1 Early/Late Integration

There has been a fair amount of debate as to the most appropriate time to integrate the audio and visual channels within the various recognition frameworks. The two most prominent integration strategies lie at opposite ends of the spectrum and are typically termed *early* and *late* integration, although hybrid strategies are rapidly growing in acceptance [142, 118, 133]. In early integration, the audio and visual feature vectors are concatenated to form one large vector, which is then used for training and recognition. In late integration, individual probabilities, or scores, are computed for each channel independently, and then the resultant probabilities are combined using some weighted, heuristic approach. In principle, if sufficient training data and an optimal learning algorithm are used, then the hybrid and late integration strategies are merely special cases of the more general early integration approach. However, practical considerations have continued to fuel the debate over the relative merits of the early and late integration approaches. These considerations include hardware that might not be capable of simultaneously acquiring the audio and visual data, and the difficulty of obtaining enough training data to adequately estimate the increased number of parameters required by early-integration architectures.

Stork et al. [128] compared the early and late integration strategies on a multi-speaker (5-speaker) recognition task of 10 consonants using time-delay neural networks. Visual data was acquired by tracking 10 reflective markers placed on the speakers' faces. Five features were extracted from the 10 pairs of (x, y) coordinates — nose-chin separation, mouth opening, mouth width, and the horizontal and vertical separation of sub-portions of the mouth. The acoustic recogniser used 14 mel-scale filter bank coefficients. They conducted audio-only, visual-only, and audio-visual recognition experiments. They termed their two audio-visual systems, "AxV" and "full AV". In their AxV system (late integration), the audio and visual recognition probabilities were computed independently and the resultant audio-visual probability was the (normalised) equally-weighted product of the two (similar to [125]). In their full AV system (early integration), the 14 audio features and 5 visual features were concatenated into a single feature vector and the neural net trained on the combined audio-visual data. Since the full AV net can learn associations between the audio and visual data at an earlier level than the AxV net, which merely treats the audio-visual

data as independent channels, it was surprising that the AxV net produced better results (error rates of 9% compared to 13%) than the full AV recogniser. The authors felt that they may not have had enough training data for the full AV net, or that it may have learnt low-level correlations between the data that were present in the training data, but not the test data.

Benoit et al. [1] also obtained superior recognition results using the late integration strategy when compared to the early integration approach, although the reasons for their findings are more clear. They investigated speaker-dependent, isolated-word experiments on a database of 54 nonsensical French words. Special blue chromatic lipstick was worn by the speaker in order to facilitate extraction of geometric measures of the lips, such as the internal and external lip width and height, inter-labial lip area, and total lip area. The acoustic signal was degraded by adding varying levels of artificial Gaussian noise. HMM word models were learnt by training on clean audio and clean visual signals. When presented with noisy acoustic data (signal-to-noise ratios of 6 dB and less) they found that the combined audio-visual recogniser performed worse than the visual-only recogniser (error rates of 32% compared to 22% at 6 dB SNR) when the early integration strategy was used. However, when a late integration strategy was used and the channel specific probabilities, $\Pr(\text{Word}|A)$ and $\Pr(\text{Word}|V)$, were combined using a weighting factor obtained from the dispersion of the four best candidates for each channel, the error rate dropped from 22% for the visual-only recogniser (82% for audio-only) to 18% for the audio-visual recogniser. Further, this late integration strategy using the weighted probabilities resulted in audio-visual recognition performance that exceeded the audio-only and visual-only performance at all noise levels. It is likely that the late integration strategy performed better than the early method primarily because the late method provided for a means of accounting for the variable level of noise in the acoustic channel, whereas the early method possessed no such capability. However, the results of their experiments do shed some light on the complexity of the audio-visual integration problem, particularly in situations where the noise level is unknown and potentially time-varying. Recognition experiments in chapter 6 shed further light on the difficulty of intelligently integrating the two channels.

2.4.2 Hybrid Integration

The difficulty of effectively integrating the audio and visual channels has led to the proposal of hybrid strategies, where clever methods are used to integrate the two channels at an intermediate stage. Sejnowski and Yuhas [123, 142, 143] proposed using a neural network to estimate the *acoustic* spectral envelope from static images of a talker. Audio-visual

recognition was accomplished by combining the spectral envelope computed directly from the (noisy) acoustic signal and the spectral envelope obtained from the neural network mapping of the image data. A weighting factor, which was empirically determined and varies linearly with the SNR of the acoustic channel, is used to optimally combine the two spectral envelopes. The resultant spectral envelope then serves as an input to a second neural network which is used for recognition. The combined audio-visual features resulted in improved recognition performance across a range of signal-to-noise ratios on a speaker-dependent, 9-vowel discrimination task.

In [118], Robert-Ribes et al. propose a similar hybrid approach, but rather than assuming that the auditory channel is dominant, they advocate projecting the audio and visual channels into an amodal motor space and then fusing the two information streams in that space. The amodal space consists of three parameters corresponding to the horizontal and vertical components of the highest point of the tongue, and the inner lip width, although it is not clear how the raw acoustic and visual data are mapped into this space. Fusion of the two channels in the motor space uses a weighted average approach similar to [142]. The authors were unable to achieve any increase in audio-visual recognition performance using this hybrid integration approach as compared to the simple late integration strategy; however, the idea of using the audio and visual channels to estimate parameters in a space representative of the complete articulatory process holds promise for further research into audio-visual integration.

Brooke and Tomlinson [133] have recently proposed an integration strategy which permits asynchrony between the audio and visual channels. Audio-visual features are obtained by concatenating the audio and visual features (the traditional early integration approach). However, for recognition, they utilise a *cross-product* HMM, where each audio-visual phone model employs a two-dimensional state transition matrix. The matrix implementation entertains the possibility of separate state transitions for the audio and visual data streams, while still providing for standard synchronised movement. In recognition experiments on connected digit triples, they have found the cross-product architecture to be superior to the more conventional left-to-right topologies across a range of signal-to-noise ratios.

The search for effective methods of integrating the audio and visual channels remains an active area of research. It is likely that further work with hybrid strategies will lead to additional insight into methods for optimally integrating the two channels. Experimental comparison of the various integration strategies was not investigated in this thesis. Rather, the early integration approach was adopted and used throughout, as it represents the most general architecture.

2.5 Discussion

It is apparent from the literature that the field of audio-visual speech recognition provides vast opportunities for research. The speechreading community has clearly demonstrated that the visual channel contains linguistically pertinent information that can be used to provide robust speech recognition of noisy speech. However, despite the successes achieved thus far, many issues remained unresolved — from how to achieve accurate lip tracking within the time constraints of this application to determining effective methods for capturing the maximal information from the two channels. Further challenges await as well, such as such as how to address the inter-speaker variability inherent in visual speech and how to extract pose-invariant visual features. Such challenges make automatic speechreading an exciting and active area of research.

This thesis focuses primarily on the real-time tracking and feature extraction problems. The lip trackers described in this thesis use a dynamic contour tracking framework which employs motion models that exploit the temporal coherence of articulating lips. Although potentially important recognition information is lost, particularly knowledge of teeth and tongue position, recognition experiments demonstrate that shape features extracted from tracking the outer lip contour are a rich source of information for audio-visual speech recognition. Further, the comprehensive shape and motion models inherent in the dynamic contour tracking framework permit the use of focused image feature detection methods. The use of more advanced feature detectors, which employ Bayesian discriminant analysis classification techniques on colour images, enables real-time tracking of unadorned lips. Further, it is shown how the dynamic contour tracking framework can be extended to the task of tracking both the inner and outer lip contours, which should enable additional reasoning to be made about the presence/absence of the teeth and tongue, and permit further benefits to be obtained from the visual channel.

3

Dynamic Contour Tracking

In order to achieve real-time tracking of the lips without resorting to expensive custom hardware, it is necessary to reduce the enormous amount of data present in images of human faces. Petajan et al. [106] attain real-time performance by using the size and position of the nostrils and the known distance between the nose and mouth to limit their search to a small window in the mouth region. In this thesis, real-time performance is achieved by using the dynamic contour tracking framework originated by Blake et al. [12, 15]. Here, Blake's tracker, which was developed to track the occluding contour of rigid, planar objects in clutter-free environments, is extended to the tracking of non-rigidly deforming, articulating lips in natural images of the face.

The power of the dynamic contour tracker framework lies in its employment of shape models, learnt dynamical motion models, and focused image feature detectors. These three components are blended together using a Kalman filter [58]. Data reduction is achieved by representing the outline of the lips as quadratic B-splines which allow smooth curvature to be modelled explicitly. Motion of the lips is assumed to be describable by a B-Spline curve parameter-set (control points) varying over time. The motion of the control points is modelled as a second order process with dynamics that imitate typical lip movements found in speech. The dynamics are used by a predictor in the tracker's Kalman filter. Measurements of the lip from the image are then combined with the predicted control point positions via the Kalman filter to provide estimates of the lip outline given by the tracker.

This chapter introduces the notation for, and gives a brief overview of, the Kalman filter-based trackers developed in this thesis. A more thorough handling of the underlying mathematics and a more detailed explanation of the intricacies of the framework can be found in [38, 37, 11, 12].

3.1 Notation

The tracker is an estimator for a moving, piecewise-smooth image-plane curve:

$$\mathbf{r}(s,t) = (x(s,t), y(s,t)).$$

The curve is represented sparsely in terms of B-splines [5], similar to the tracking work of others [97, 30]. The lip outline is parameterised by Quadratic splines ($d = 3$) of length L with multiple knots for the vertices (lip corners)

$$x(s) = \mathbf{B}(s)\mathbf{X} \ \text{ and } \ y(s) = \mathbf{B}(s)\mathbf{Y}, \ \ 0 \leq s \leq L \tag{3.1}$$

where $\mathbf{X} = (X_1, .., X_{N_X})^T$ and similarly for $\mathbf{Y}$ with $N_X = L$ for closed curves and $N_X = L + d - 1$ for open ones (with appropriate variations where multiple knots are used to vary curve continuity). The elements of $\mathbf{X}$ and $\mathbf{Y}$ are simply the x and y coordinates corresponding to the control points (X_m, Y_m) of the B-spline. The vector $\mathbf{B}(s)$ maps the control point vector $\mathbf{X}$ to its associated curve $x(s)$.

For work using parametric spline curves, it is useful to define a norm in order to measure how closely one curve approximates another. In this work, the L_2-norm $\|\mathbf{X}\|$ in *spline* space is equivalent to the root mean square distance in the image plane where

$$\|\mathbf{X}\|^2 = \int_{s=0}^{L} x(s)^2 ds$$

which can be written as

$$\|\mathbf{X}\|^2 = \mathbf{X}^T \mathcal{H}_o \mathbf{X} \tag{3.2}$$

with

$$\mathcal{H}_o = \int_0^L \mathbf{B}^T(s)\mathbf{B}(s)ds. \tag{3.3}$$

Given this norm, the inner product can be similarly defined as

$$\langle \mathbf{X}, \mathbf{X}' \rangle = \mathbf{X}^T \mathcal{H}_o \mathbf{X}', \tag{3.4}$$

which enables extraction of individual shape parameters from splines.

3.2 Tracking

The tracking problem is to estimate the lip motion which is assumed to be describable by a B-spline of predefined form with control points $(\mathbf{X}(t), \mathbf{Y}(t))$ varying over time. The tracker generates *estimates* of these control points, denoted $(\hat{\mathbf{X}}(t), \hat{\mathbf{Y}}(t))$, and the aim is that these estimates should represent a curve that, at each time-step, matches the lip outline as closely as possible. The tracker consists, in accordance with standard practice in temporal filtering [54, 4], of two parts: a system model and a measurement model. Broadly speaking, the system model specifies the likely dynamics of the curve over time, relative to an average shape or "template" [49] whose control points are given by $(\overline{\mathbf{X}}, \overline{\mathbf{Y}})$, and which is generated by an interactive drawing tool, while the measurement model specifies the positions along the curve at which measurements are made and how reliable they are.

The lip template was created by fitting a spline to a set of closed lips. Its outline is similar to the one shown in figure 3.1.
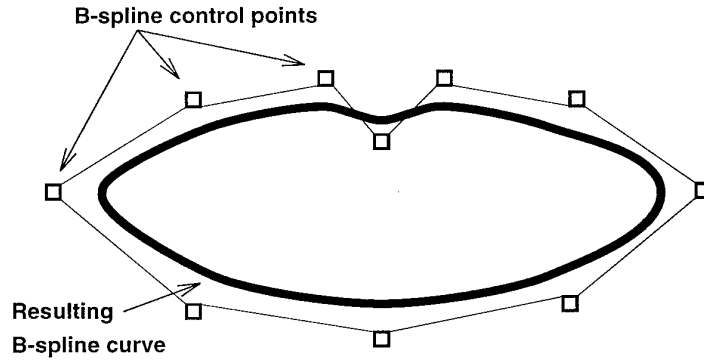
Figure 3.1: *Lip template, $(\overline{\mathbf{X}}, \overline{\mathbf{Y}})$, showing control points (boxes) and B-spline fit to a set of closed lips.*

3.3 Reduced Tracking Space

The tracker could conceivably be designed to allow arbitrary variations in control point positions over time. This would allow maximum flexibility in deforming to the various lip shapes; however, this is known to lead to instability in tracking [12] when following complex shapes which require many control points. Thus, the number of degrees of freedom was limited by imposing shape constraints on the deforming contour. Deformations of the contour, represented as a control point vector $(\mathbf{X}, \mathbf{Y})$ in *spline* space $\mathcal{S}_X$, are restricted to lie in a *shape* space $\mathcal{S}_Q$ represented by a shape-vector $\mathbf{Q}$. Transformations between

control point vectors $(\mathbf{X}, \mathbf{Y})$ and shape-vectors $\mathbf{Q}$ are made by projecting the spline onto the shape-matrix W using a least-squares fit

$$\mathbf{Q} = W^{\dagger} \begin{pmatrix} \mathbf{X} - \overline{\mathbf{X}} \\ \mathbf{Y} - \overline{\mathbf{Y}} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = W\mathbf{Q} + \begin{pmatrix} \overline{\mathbf{X}} \\ \overline{\mathbf{Y}} \end{pmatrix} \tag{3.5}$$

where $W^{\dagger}$ is a pseudo-inverse defined by

$$W^{\dagger} = (W^T \mathcal{H} W)^{-1} W^T \mathcal{H} \tag{3.6}$$

with $\mathcal{H}$ consisting of the sub-matrices, $\mathcal{H}_o$, defined in (3.3)

$$\mathcal{H} = \begin{pmatrix} \mathcal{H}_o & \mathbf{0} \\ \mathbf{0} & \mathcal{H}_o \end{pmatrix} \tag{3.7}$$

and W defined below.

Blake and Curwen [12] showed the need for accurate shape models in contour tracking to provide stability to the tracker. Their application was the tracking of hands which were treated as rigid, planar objects. Since it is known that under orthographic projection a rigid, planar shape has only 6 degrees of freedom — the parameters of an affine transformation [134, 81] — they successfully tracked rigid hand motion by limiting the deformations to a 6-dimensional affine space. While the lips are neither rigid nor planar, it turns out that the general symmetry of the lips results in motions that can be roughly approximated by affine deformations of a 2D lip template. A basis, W, for this affine space is

$$W = \left\{ \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}, \begin{bmatrix} \overline{\mathbf{X}} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \overline{\mathbf{Y}} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \overline{\mathbf{X}} \end{bmatrix}, \begin{bmatrix} \overline{\mathbf{Y}} \\ \mathbf{0} \end{bmatrix} \right\} \tag{3.8}$$

where N_X-vectors $\mathbf{0}$ and $\mathbf{1}$ are defined by:

$$\mathbf{0} = (0, 0, .., 0)^T \quad \mathbf{1} = (1, 1, .., 1)^T. \tag{3.9}$$

These affine deformations which represent lip movements in terms of horizontal and vertical translation, scaling, and shearing of the lip template are shown in figure 3.2.

In tracking tests it was found that the affine deformations accounted for 91% of the overall lip motion variance; however, in order to track the more subtle lip movements, it was necessary to permit non-affine motion. These additional lip motions were incorporated into the model by choosing a shape space capable of representing both affine and non-affine deformations. The shape-matrix W was extended by one vector for each permitted non-affine degree of freedom. These additional lip motions were derived from "key-frames" — representative non-rigid deformations of the template formed by fitting splines to expressions
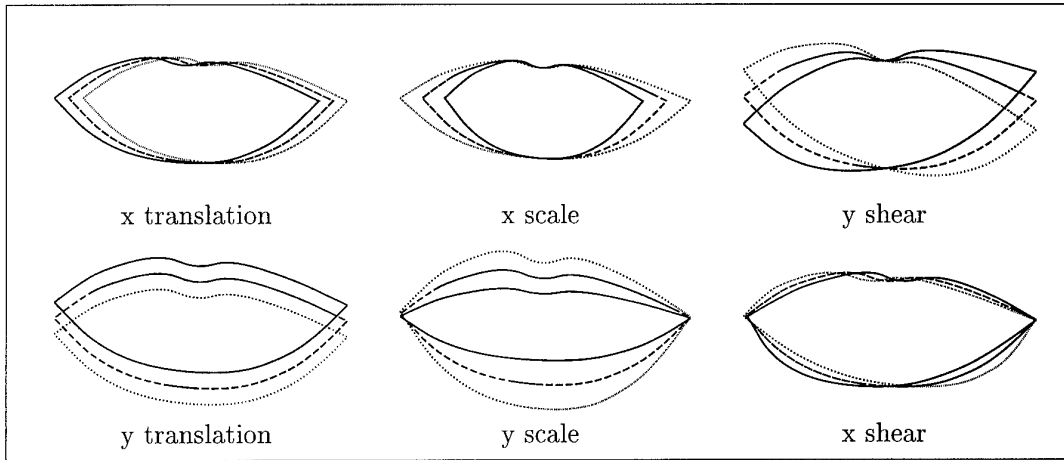
Figure 3.2: *Lip movements corresponding to* **affine** *deformations of the mouth template plotted a normalised displacement either side of their mean. The first two components represent horizontal and vertical displacement/translation. The third and fourth, horizontal and vertical scaling, and the fifth and sixth, vertical and horizontal shearing.*

such as 'ah', 'ee', and 'oo'. Figure 3.3 shows typical key-frames used in tracking the lips from the frontal view.

The complete tracking space is thus spanned by the six affine deformations plus the additional key-frame deformations. The resulting basis expressed in terms of the original template $(\overline{\mathbf{X}}, \overline{\mathbf{Y}})$ and the key-frames $(\mathbf{X}_1, \mathbf{Y}_1), \cdots (\mathbf{X}_{N_k}, \mathbf{Y}_{N_k})$ is

$$W = \left\{ \underbrace{\begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}, \begin{bmatrix} \overline{\mathbf{X}} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \overline{\mathbf{Y}} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \overline{\mathbf{X}} \end{bmatrix}, \begin{bmatrix} \overline{\mathbf{Y}} \\ \mathbf{0} \end{bmatrix}}_{\text{affine basis}}, \underbrace{\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{Y}_1 \end{bmatrix}, \cdots, \begin{bmatrix} \mathbf{X}_{N_k} \\ \mathbf{Y}_{N_k} \end{bmatrix}}_{\text{key-frames}} \right\}. \quad (3.10)$$

Although the construction of shape spaces using the key-frame building approach results in basis vectors corresponding directly to known lip deformations, often times the resultant tracking spaces are unnecessarily large. Tracking spaces possessing more degrees of freedom than are necessary are undesirable for two reasons. First, the computational cost of the tracking algorithm is $\mathcal{O}(N_Q^3)$, where N_Q is the dimension of the shape space, so the computational penalty for employing unnecessarily large spaces can be severe. The second reason is that additional, non-essential degrees of freedom can lead to tracking instabilities.

The most natural method for constructing shape spaces is to *learn* the space of lip deformations from training data of sample motions. The strength of this approach is that tracking spaces can be customised to the visual speech patterns of individual talkers. Having obtained training data consisting of prototypical lip movements linked to speech production for a particular speaker, principal components analysis (PCA) [31] provides an efficient
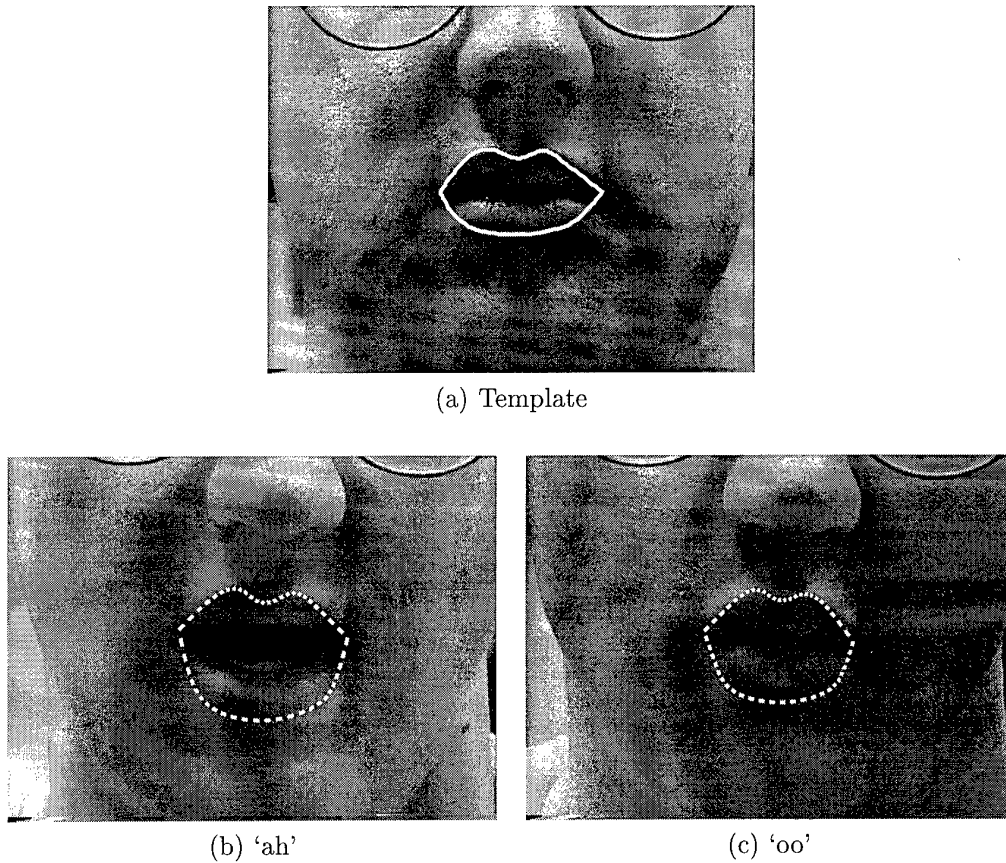
(a) Template

(b) 'ah'

(c) 'oo'

Figure 3.3: *Additional degrees of freedom are permitted by adding key-frame deformations such as these to the tracking space.*

means for capturing the principal modes of lip variation. Further, the shape deformations resulting from the PCA can be ordered according to the percentage of overall lip motion that they can describe. This provides a convenient method for determining the size of the shape space to use for tracking, which can be expressed in terms of the percentage of variance accounted for by the basis vectors chosen.

3.4 Predictive Dynamics

The motion of the lips was modelled as a second order process driven by noise, which is widely used in control theory [2]. The choice of a second order model permitted constant velocity motion, decay, and damped oscillation. A state-space representation is used with the state vector $\mathcal{X}_n$ defined in terms of the shape vector $\mathbf{Q}$

$$\mathcal{X}_n = \begin{pmatrix} \mathbf{Q}_n \\ \dot{\mathbf{Q}}_n \end{pmatrix}. \tag{3.11}$$

The system can be described in discrete time by the difference equation

$$\mathcal{X}_{n+1} - \overline{\mathcal{X}} = A(\mathcal{X}_n - \overline{\mathcal{X}}) + \begin{pmatrix} \mathbf{0} \\ B\mathbf{w}_n \end{pmatrix}. \tag{3.12}$$

Here A is a $2N_Q \times 2N_Q$ matrix defining the deterministic part of the dynamics and

$$\overline{\mathcal{X}} = \begin{pmatrix} \overline{\mathbf{Q}} \\ \overline{\mathbf{Q}} \end{pmatrix}.$$

The driving noise $\mathbf{w}_n$ is white with shaping matrix B. Without loss of generality, the matrix A can be expressed as

$$A = \begin{pmatrix} 0 & I \\ A_0 & A_1 \end{pmatrix}$$

permitting the equations of motion to be simplified to the standard linear predictor form:

$$\mathbf{Q}_{n+2} = A_0\mathbf{Q}_n + A_1\mathbf{Q}_{n+1} + (I - A_0 - A_1)\overline{\mathbf{Q}} + B\mathbf{w}_n. \tag{3.13}$$

A's eigenvectors represent modes of oscillatory motion, and the corresponding eigenvalues give natural frequencies and damping constants for those modes. Default dynamics for A are initially set manually, but are ultimately learnt using tracked sequences of lip deformations as described in [14, 15].

3.5 Measurement Model

Measurements of the lip contour position are made by searching along normals, $\hat{\mathbf{n}}(s,t)$, to the predicted lip position, $\tilde{\mathbf{r}}(s,t) = A(\mathcal{X}_n - \overline{\mathcal{X}})$, in the image for features. These features correspond to the boundary of the object being tracked, eg. the lips. In the simplest case, features are obtained by applying one-dimensional image operators, such as edge detectors, to the grey-level intensities along the normals. However, more advanced feature detectors, which are required when tracking poorly contrasted boundaries such as between the lips and facial skin, permit the matching of statistical templates, or alternately, employ Bayesian classification techniques to colour image data, in order to identify the object boundary.

This can be described more formally by

$$\nu(s,t) = [\mathbf{r}(s,t) - \tilde{\mathbf{r}}(s,t)].\hat{\mathbf{n}}(s,t) + v(s,t) \tag{3.14}$$

where $\nu(s,t)$ represents the displacement of the image feature relative to the predicted curve $\tilde{\mathbf{r}}(s,t)$, and $v(s,t)$ is the spatial measurement noise, assumed Gaussian with zero mean and covariance R_s that varies with position, s, but is taken to be temporally constant. The innovations, $\nu(s,t)$, are defined only along normals to the curve as the tangential motion

is unobservable locally — the well known aperture problem [68]. The spatial measurement covariance R_s is a function of several variables including the electrical noise involved in image formation, spatial camera noise, and the detection of erroneous features obtained via the feature detection process. If heavy background clutter is present, then the resultant false features often lead to non-Gaussian measurement probability densities which require more sophisticated modelling [71]; however, in practice, if only moderate clutter is present and prudent feature detection methods are used, it is often possible to approximate the measurement densities as Gaussians.

The feature measurements, or innovations, $\nu(s,t)$ are related to the state vector $\mathcal{X}$ by the observation matrix $H(s,t)$ given by

$$\nu(s,t) = H(s,t)(\mathcal{X} - \tilde{\mathcal{X}}) + v(s,t) \tag{3.15}$$

where from (3.1), (3.5), and (3.14),

$$H(s,t) = \left[\ \left(\ \hat{\mathbf{n}}_x(s,t)\mathbf{B}(s) \quad \hat{\mathbf{n}}_y(s,t)\mathbf{B}(s) \ \right) W \quad \mathbf{0} \ \right]. \tag{3.16}$$

In theory $H(s,t)$ is a continuous function of s, although in practice the curve is not observed in its entirety but rather at sampled points, s_i, along the contour. The normals to the contour at these sampled points are referred to as *search lines* and are shown in figure 3.4 for the frontal lip tracker.

As the next chapter provides methods for identifying the boundary between the lips and skin along the various search lines, it is appropriate to introduce some extra notation here. The grey-level or colour (red-green-blue) intensity of a pixel a signed-distance r along a normal to the contour $\hat{\mathbf{n}}(s_i,t)$ will be denoted as $I_i(r)$, where i represents the number of the search line. For example, the grey-level intensity profile along the normal to the middle of the bottom lip (search line 23 in figure 3.4) is shown in figure 3.5. Thus the intensity $I_i(r)$ can be thought of as the intensity of the pixel at location $\mathbf{r}(s_i,t) + r\hat{\mathbf{n}}(s_i,t)$. Various operators, or feature detectors, are applied to the intensity profiles $I_i(r)$, with the goal being identification of a "feature" at the lip-skin boundary.

3.6 Kalman Filter

Tracking is accomplished using a standard Kalman filter [54, 58] consisting of prediction and measurement assimilation steps. Since the distribution of the state estimates $\hat{\mathcal{X}}_n$ is assumed to be Gaussian, all of the information about the lip position is carried by the 1st order (mean) and 2nd order (covariance) moments. For each time step, from $t - \Delta$ to t,
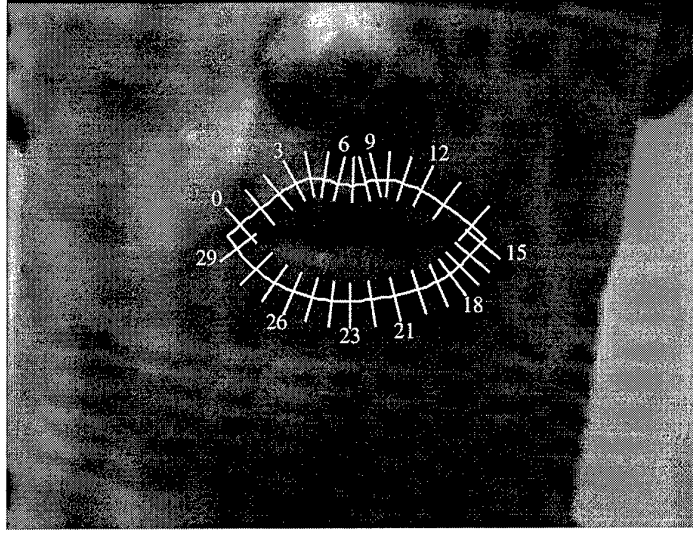
Figure 3.4: *Measurements are taken along normals at sampled positions of the lip contour. These search lines are labelled clockwise from 0 − 29. Image features are identified through image processing operations on the search line data and the resultant innovations are incorporated into the Kalman filter.*
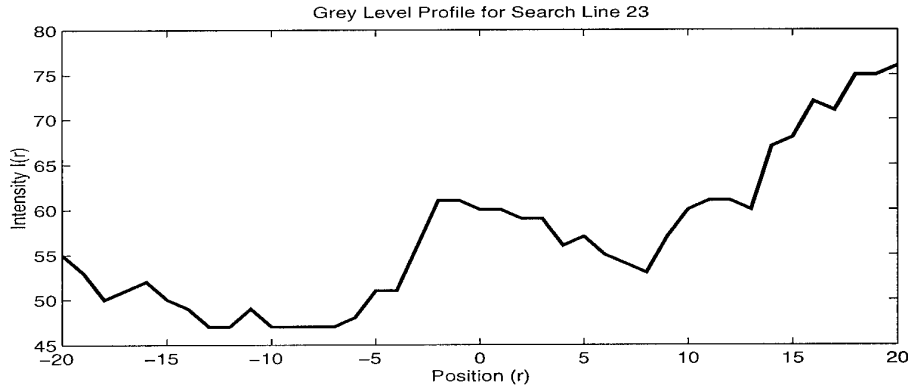

Figure 3.5: *Grey-level intensity profile along the lower lip. The lip-skin boundary occurs at r = 0, while r < 0 represents the intensity distribution of the skin, and r > 0 corresponds to the lips.*

prediction consists of the deterministic drift due to the system dynamics

$$\tilde{\mathcal{X}}_n = A\mathcal{X}_{n-1} + (I - A)\overline{\mathcal{X}} \tag{3.17}$$

and diffusion due to the driving noise

$$P_n = AP_{n-1}A^T + \begin{pmatrix} 0 & 0 \\ 0 & BB^T \end{pmatrix} \tag{3.18}$$

where P_n is the covariance of the estimate $\hat{\mathcal{X}}$ at time $t = n\Delta$.

The state uncertainty, P_n, has been found to be valuable not only for calculation of the Kalman gain K, but also in limiting the extent of the feature search along the search lines in the form of a *validation gate* [4]. For a given search line i, the correct image feature should fall within a range determined by the uncertainty in the state estimate and the measurement uncertainty, R_{s_i}. The covariance associated with the positional uncertainty of the ith feature, $\Omega(s_i)$, is given by

$$\Omega(s_i) = H(s_i)PH^T(s_i) + R_{s_i}. \tag{3.19}$$

In general the covariance of the sensor error R_{s_i} represents a two dimensional uncertainty ellipse; however, owing to the nature of the feature search mechanism, that is, image feature measurements are made only along normals to the curve, tangential displacement is unobservable. Thus, if the one-dimensional measurement covariance for the feature detection process is denoted as $\sigma_{s_i}^2$, then the covariance will be $\sigma_{s_i}^2$ along the normal and infinite tangential to the normal. The resultant 2D measurement covariance, R_{s_i}, is no longer defined, although its inverse can be found using the 1D measurement error,

$$R_{s_i}^{-1} = \sigma_{s_i}^{-2}\hat{\mathbf{n}}(s_i)\hat{\mathbf{n}}^T(s_i). \tag{3.20}$$

At each search position s_i along the curve, the image feature will fall within γ standard deviations of the predicted position with the associated degree of certainty

$$\nu^T(s_i)\Omega^{-1}(s_i)\nu(s_i) \le \gamma^2. \tag{3.21}$$

This innovation uncertainty can then be used to determine the validation gate, or required search scale, along the search line normal. That is, the search scale for the ith feature, ρ_i, along its normal, $\hat{\mathbf{n}}(s_i)$, can be found from (3.21) by

$$\rho_i = \sqrt{\frac{\gamma^2}{\hat{\mathbf{n}}^T(s_i)\Omega^{-1}(s_i)\hat{\mathbf{n}}(s_i)}}. \tag{3.22}$$

Following the prediction step, measurements are made in the image as previously described. If the measurements along each of the search lines are assumed to be mutually independent, then the measured image features can be iteratively assimilated into the curve estimate at each sampled position

$$\mathcal{X}_n = \tilde{\mathcal{X}}_n + K_{s_i}\nu(s_i) \tag{3.23}$$

where $\nu(s_i)$ is the innovation along the normal at s_i as before, and the Kalman gain, K_{s_i}, for each measurement is given by

$$K_{s_i} = P_n H^T \left(HP_n H^T + R_{s_i}\right)^{-1}. \tag{3.24}$$

The reactive effect of the measurements then decreases the uncertainty in the estimated state which is updated accordingly

$$P_n = P_n - K_{s_i} H P_n. \tag{3.25}$$

3.7 Learning Model Dynamics

As is the case with any control system, it is important to know the actual parameters of the plant or process to be modelled. Typically the plant dynamics are determined from the underlying physics of the process [44, 73]. Several researchers [132, 85, 59] have developed detailed physiological models of the facial muscles that have been useful for describing facial expressions and lip movement. However, there exists an alternative to such detailed modelling, which is to learn the dynamics of moving lips from actual sequences of connected speech. Naturally there is the chicken and the egg problem, where in order to learn lip motions from a tracked sequence, it is first necessary to be able to track the lips. To combat this, default dynamics are set by hand which permit lip tracking of slow-speech. The output of this default tracker is then used, via a learning algorithm [15], to generate a new tracker with improved dynamics. This new tracker, which is considerably more stable than the default tracker, as it has been tuned to follow only speech-like lip movements, is then used to track normal speech from which newer dynamics can be learnt. This bootstrap tracking-learning process can be iteratively repeated until a desired level of tracker tuning is achieved; however, in practice, two iterations of this training procedure have been sufficient.

The principal strength of this approach is that the motion models can be finely tuned to the visual articulatory patterns of the speaker. There is, however, a disadvantage to this learning method in that the tuned tracker is best suited for motions seen in the training sequence and may have difficulty tracking movements not yet observed. For instance, if during training the speaker makes head-nodding movements but little or no horizontal head movements, the tuned tracker will learn that horizontal movements are to be damped out and may not be able to track in the presence of horizontal head movements. In an attempt to deal with such problems, investigation into the coupling of head and lip trackers is currently in progress [117, 78]. In this thesis, which focuses primarily on movements associated with articulating lips rather than on global head movements, the speakers spoke naturally, limiting unnecessary head movements. Training sequences typically consisted of phonetically balanced Central Institute for the Deaf (CID) Everyday Speech sentences [66] or similar continuous speech containing all of the phonemes.

3.8 Summary

This chapter has provided an overview of the dynamic contour tracking framework. This framework serves as the foundation for many of the ideas developed and presented in this thesis. The key to the success of the tracking framework lies in its utilisation and integration of three powerful techniques:

- Prior shape models

- Learnt motion dynamics

- Focused image feature detectors

The combination of these techniques, in conjunction with the sparse representation of the lip outline using a B-Spline parameterisation, enables efficient utilisation of the computational resources. This results in Kalman filter-based trackers capable of operating at real-time rates without recourse to special hardware.

In the next chapter we will see that, due to the poor contrast in the mouth region, it is difficult to obtain reliable image feature measurements which accurately identify the boundary between the lips and facial skin. Consequently, several different feature detection methods are investigated and presented.

4

Lip Tracking

Depth discontinuities between the occluding contours of objects and their surroundings often result in readily identifiable boundaries or edges. This reality has led to the development of fast and reliable edge detection algorithms [28, 41, 61]. Consequently, edge detection methods are widely used in tracking applications [75, 38, 60]. However, when tracking objects in natural scenes or smooth objects with little depth discontinuity, such as the articulatory movements of deforming lips, edge detection schemes have proven to be inadequate [34, 120, 122]. The primary reason for this is that the lips are set against flesh-tones with consequently weak contrast [144]. In addition, spurious edges (clutter) are often present that are not due to the boundary between the lips and the surrounding skin, but rather the texture of the face and lips. This dual problem of a lack of features on one hand and an over abundance of *false* features on the other is particularly troublesome for contour-based trackers where only *one* measurement for each search line can be integrated into the Kalman filter (section 3.6). Methods exist for combining multiple observed features using a probabilistic data association filter (PDAF) [4, 114] by weighting each observation with the probability that it originated from the target. However, the PDAF assumes that the target detection probability distributions and the probabilities of obtaining false measurements are known, which is often not the case. Further, the PDAF does not solve the more fundamental problem which is that there simply are not edges along the lower lip boundary. This point is clearly demonstrated in figure 4.1 where there is both an absence

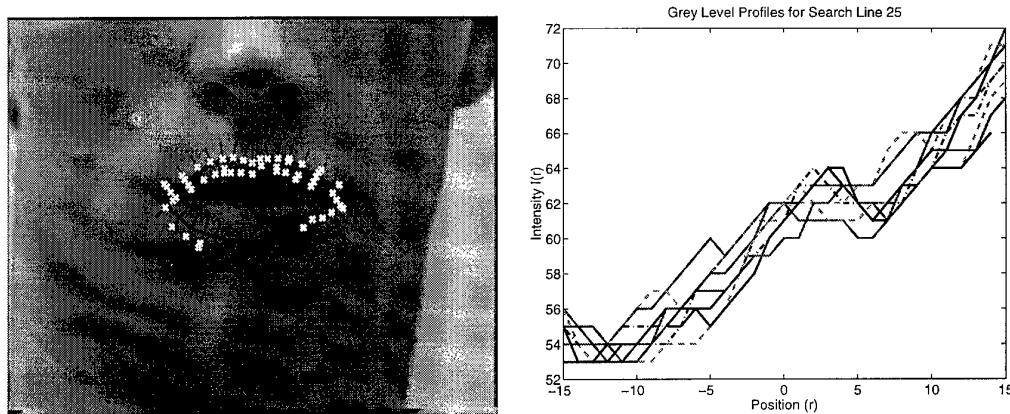of edges along the lower lip and distracting edges along the upper lip.

Figure 4.1: *There is little contrast between the boundary of the lower lip and the surrounding skin. The white crosses represent the edges found on the search normals (thin black lines). Note the absence of edges around the lower lip and the presence of distracting edges along the top lip due its texture. The graph on the right shows typical intensity profiles for 1D search lines along the lower lip. The absence of a sharp jump or sudden change in the intensity profiles similarly demonstrates the difficulty in delineating the lip boundary. Note also that the total variation in intensity along the profile is less than 20 grey-levels. The actual location of the lower lip boundary corresponds to the position $r = 0$.*

As discussed in section 3.5, accurate and reliable feature measurements are a necessary ingredient in the dynamic contour tracking framework. Thus, in addition to developing precise shape and motion models, there remains the difficult problem of how best to extract image *features* which can accurately delineate the lip outline.

. This chapter explores various solutions to the problem of tracking unadorned lips within the real-time constraints imposed by the audio-visual speech recognition application. A disciplined approach to the lip tracking problem is taken. First, the tracking of the lips from the profile view is considered, where the sharply silhouetted mouth and limited articulatory movements simplify the problem. Next, tracking from frontal views is investigated with the assistance of lipstick to enhance the contrast of the lips. This tracker enables the development of customised shape and motion models which are then utilised in unadorned-lip trackers which employ advanced feature detection techniques to identify the lip boundary.

Several different methods for identifying the boundary between the lips and surrounding skin are devised and their discriminating potential assessed. A data-driven approach, that is, the use of statistical models of the grey-level appearance of the mouth region, is shown to be the most successful.

4.1   Profile Lip Tracking

When viewed from the profile, the mouth appears sharply silhouetted against its background, resulting in easily identifiable features. Further, the articulatory movements of the lips appear less complex in profile viewing. For these reasons tracking the lip profile is favourable to tracking from the frontal view. Thus, in order to test the feasibility of using the dynamic contour framework for tracking the non-rigidly deforming lips, initial tracking used the profile view.

The tracking space was constructed using the key-frame building approach (section 3.3) and consisted of three affine components (X and Y translation and isotropic scaling) plus two non-rigid deformations corresponding to lip puckering and curling of the lower lip. The system dynamics and driving noise were learnt using the bootstrap learning procedure discussed in section 3.7 and [15]. Following the learning process, lip tracking was stable and sufficiently agile to follow normal speech, including plosives which can be particularly rapid. This can be seen in figure 4.2 where a tracked sequence of the word "four" is shown from this view.

The profile lip tracker proved to be accurate and stable, and simple speech recognition experiments were conducted using it (section 6.2.1) which demonstrated the benefit of incorporating visual information into acoustic-only speech recognisers. However, it is known that human lipreaders rely on information about the presence/absence of the teeth and the tongue [25, 94, 131]. Thus, from a speech recognition standpoint there is a potential loss of information in profile viewing in that the tongue and teeth are no longer visible. There may also be a loss of shape information in the lip contour itself, since its width is no longer directly observable in profile. Our experiments (chapter 6 and [39, 77]) and those of others [9] suggest that lip width is important for discriminating words in audio-visual speech recognition. In addition, Benoit et al. [9] evaluated recognition performance on nonsensical French words using parameters extracted from frontal and profile views. In their experiments, error rates for profile views were twice that attained using frontal views (40% versus 20%). For these reasons subsequent efforts address the frontal lip tracking problem.

4.2   Inner Lip Contour Tracking

Experience had shown that there were insufficient edge features around the outer lip contour to accurately define the lip boundary, so tracking of the inner lip contour (inner mouth) was investigated as an alternative. Promising work by Moses et al. [100] showed that there
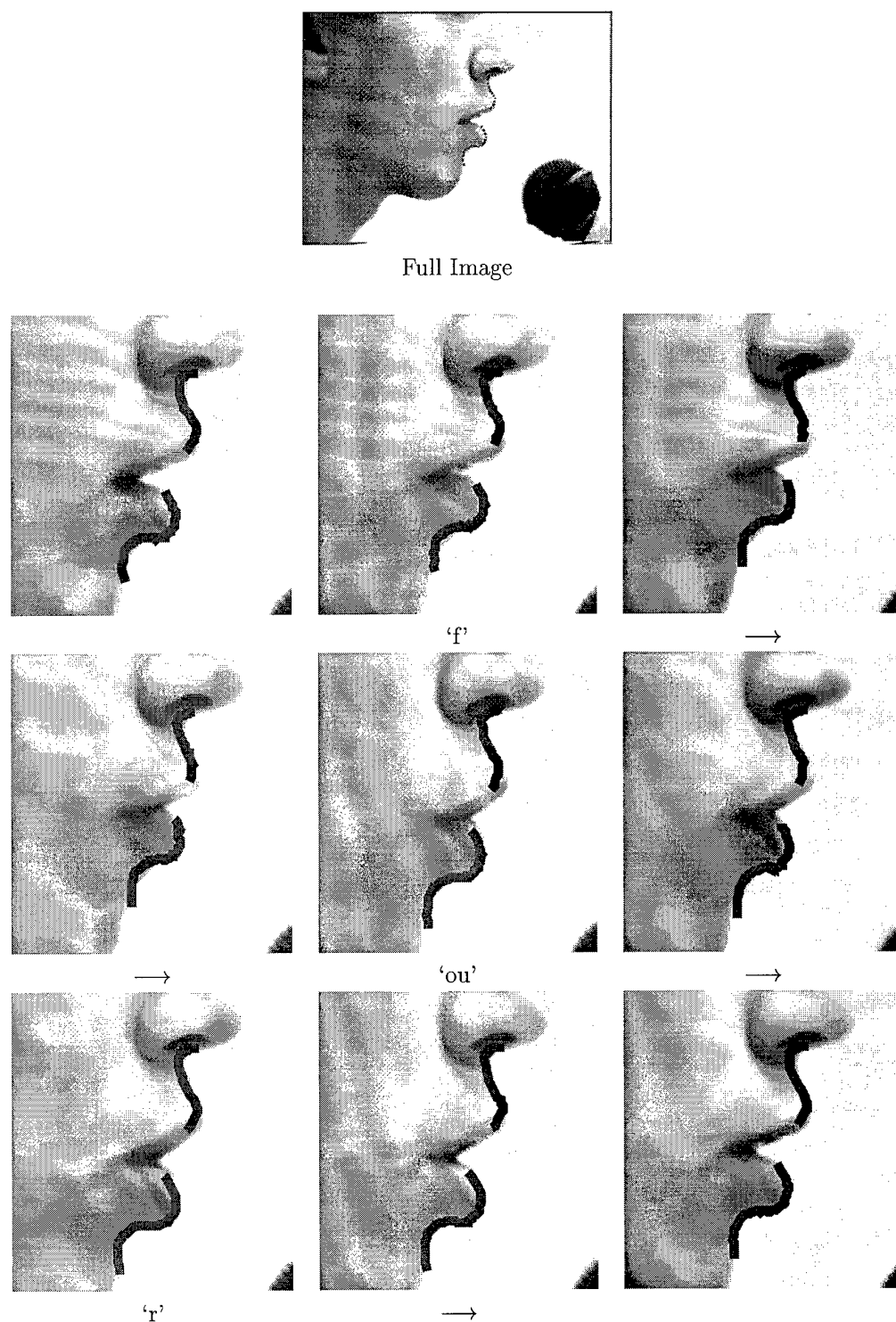
Full Image

'f'

$\longrightarrow$

$\longrightarrow$ 'ou' $\longrightarrow$

'r' $\longrightarrow$

Figure 4.2: *Tracking the word "four". Snapshots taken approximately every 40 ms. The tracker accurately follows the lower lip during the f-tuck (curling of the lip to form the 'f' sound) in tracked frames 3 and 4 and continues tracking through the lowering of the jaw necessary for the 'our' sound.*

existed a grey-level intensity valley between the lips which was invariant to illumination, viewpoint, identity, and expression. Using this fact, Reynard et al. [117, 100] developed a prototype inner-lip contour tracker where this intensity valley was used to locate the corners of the mouth and the inside of the upper lip. Standard edge features (pixel gradients) were used to identify the inside of the lower lip.

'f'   'i' $\longrightarrow$

'v'
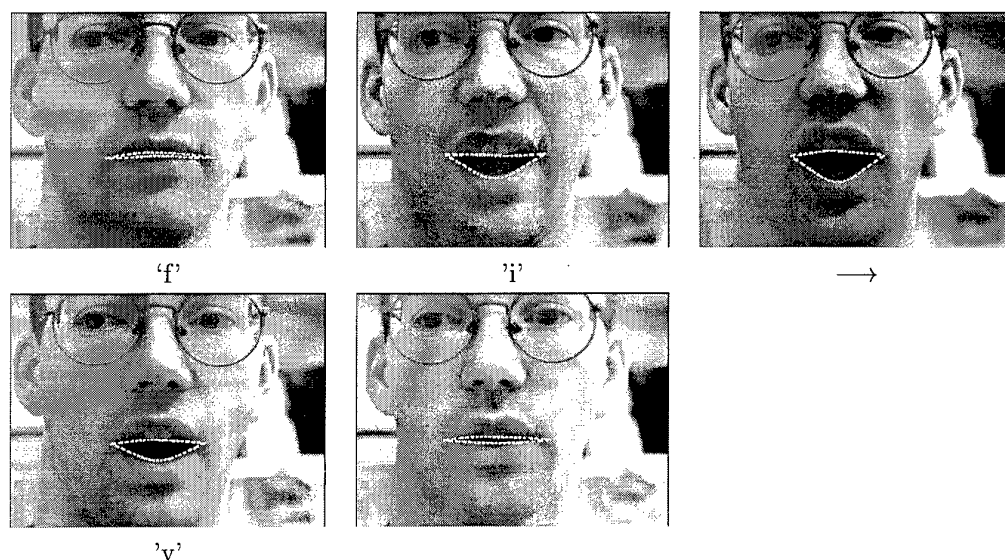
Figure 4.3: *Inner contour tracking of the word "five". Snapshots taken approximately every 60 ms. Whilst tracking is stable and the outline closely approximates the inner mouth region, the upper lip contour becomes confused by the presence of the teeth mistaking them for the inner lip and continues to track them throughout the sequence.*

Some success was achieved using this tracker, although several problems surfaced, some due to distracting features within the mouth and others due to degenerate conditions caused by closed mouth conditions. This can be seen in figure 4.3 where a tracked sequence of the word "five" is shown. Whilst the tracker reliably approximates the inner mouth region, some problems are evident. First, the upper tracked contour has an affinity both for the inside lip and for the teeth when visible, whereas clear differentiation of lips and teeth is a requirement for reliable speech recognition. Secondly, whilst tracking is stable during speech, lateral head-motions can cause errors when lock is lost on the mouth corners. This is amplified by the fact that measurements yield only one dimensional information normal to the curve when the mouth is nearly closed (flat contour). Thus, no information is available about the horizontal positioning of the mouth which often leads to instabilities during tracking. However, even in the absence of lateral motion, it is difficult to pinpoint the mouth corners accurately — the dark visual feature (valley) tends to extend beyond the

mouth, resulting in the slightly elongated contour. This is particularly troublesome because it is known from visual speech recognition experiments [9, 77] that the width of the mouth (oral cavity) contains important recognition information for word discrimination tasks. In addition, although the term "corner" might suggest the use of a classic corner detector [61], the expression is misleading as the mouth corner is not a "corner" in the conventional sense, but rather the end of the valley, which is notoriously difficult to accurately pinpoint. Lastly, avoiding and/or recovering from degenerate conditions such as a closed mouth remains a problem, because as the oral cavity closes, the feature search mechanism permits edges from the top lip to become confused with those from the bottom, and vice versa, which can result in the tracker inverting.

These problems can be overcome. For example, Petajan et al. [107] use the intensity valley along with colour thresholding to find the inner mouth region. However, they overcome the aperture problem by using the location of the nostrils to locate the mouth region and they avoid the contour degeneracy problem by using a region-growing approach as opposed to actually tracking the inner lip contour. The tracker could be customised to the task of tracking the inner lip contour by building special purpose detectors to accurately identify the end of the valley at the lip corners. Additional customising could be used to limit the search range along the inner lip contour to prevent the top lip from becoming confused with the bottom lip and also to avoid degenerate conditions. However, it was decided to develop more general feature search mechanisms with applicability beyond lip tracking. So, despite the promising aspects of tracking the valley and the inner lip contour, tracking of the outer lip contour is preferable to the inner lip contour.

4.3 Cosmetically-Assisted Outer Lip Contour Tracking

The complex motions of the non-rigidly deforming lips necessitates the use of shape and motion models tuned to the articulatory patterns of the speaker. In order to obtain these models it was first necessary to be able to track the outer lip contour, and therefore lipstick was used to enhance the contrast around the lips. This resulted in clearly identifiable edges at the lip boundary which enabled tracking using large ($\geq$ 10 dimensional) shape spaces and default dynamics set by hand. This tracker was then used to gather sequences of lip deformations from which the motion dynamics were learnt. A tracker employing the learnt dynamics was then used to gather additional sequences. Principal components analysis (described later in section 6.1.2) was then performed on the tracked sequence in order to obtain a computationally tractable six-dimensional shape space which accounted

for 99% of the overall variance of the lip motion. The resultant tracker possessed a shape space learnt from actual deformations of the speaker's lips, and, in addition, motion models which captured the temporal coherence of the speaker's lip movements. This tracker was stable, robust, and very accurate. This is corroborated by figure 4.4, where the contour follows the rapidly moving lip outline throughout an utterance of the word "previous".
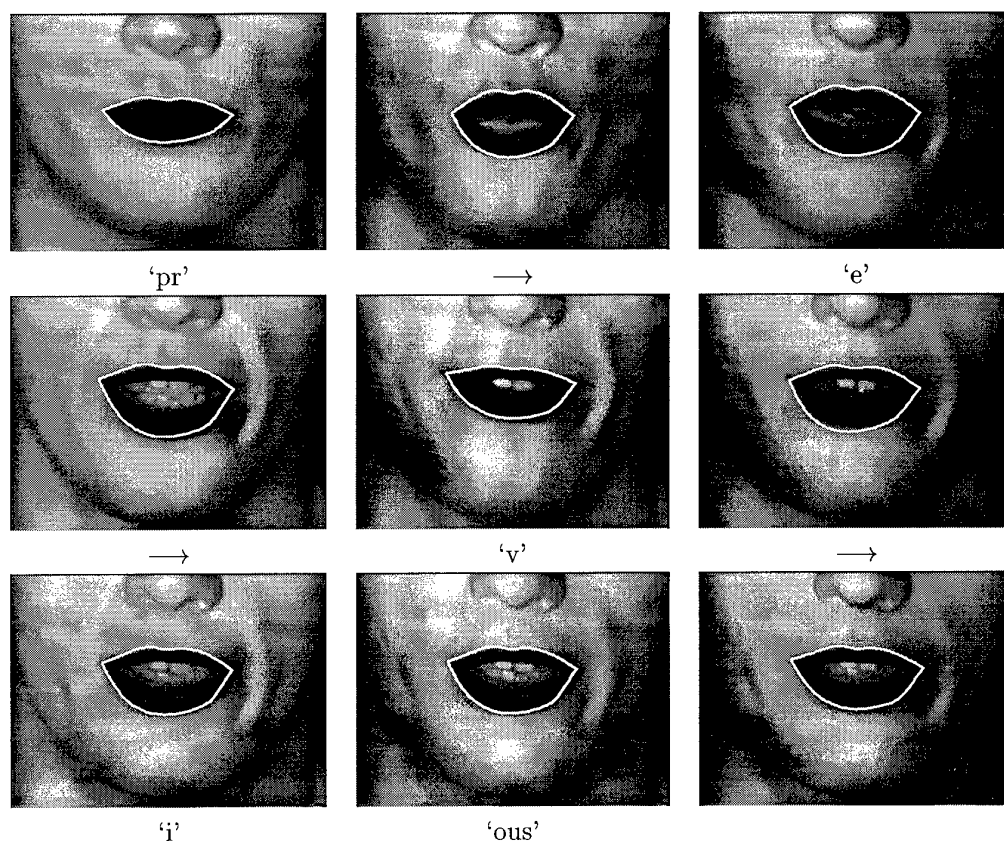

Figure 4.4: *Tracking the word "previous". Snapshots taken approximately every 80 ms. The white line represents the position of the contour after the measurements have been assimilated. The fully trained tracker accurately follows the deforming lips during the entire sequence including the protrusion (frame 2) and the horizontal spreading of the 'e' (frame 4). It also tracks the rapidly moving plosive 'p' and the labio-dental 'v' without difficultly.*

Further, tracking is also robust to changes in head position and pose. For example, in figure 4.5 the speaker simultaneously nods his head while saying "six". The tracker accurately follows the lips throughout the entire sequence, including when the lips and head move concurrently.
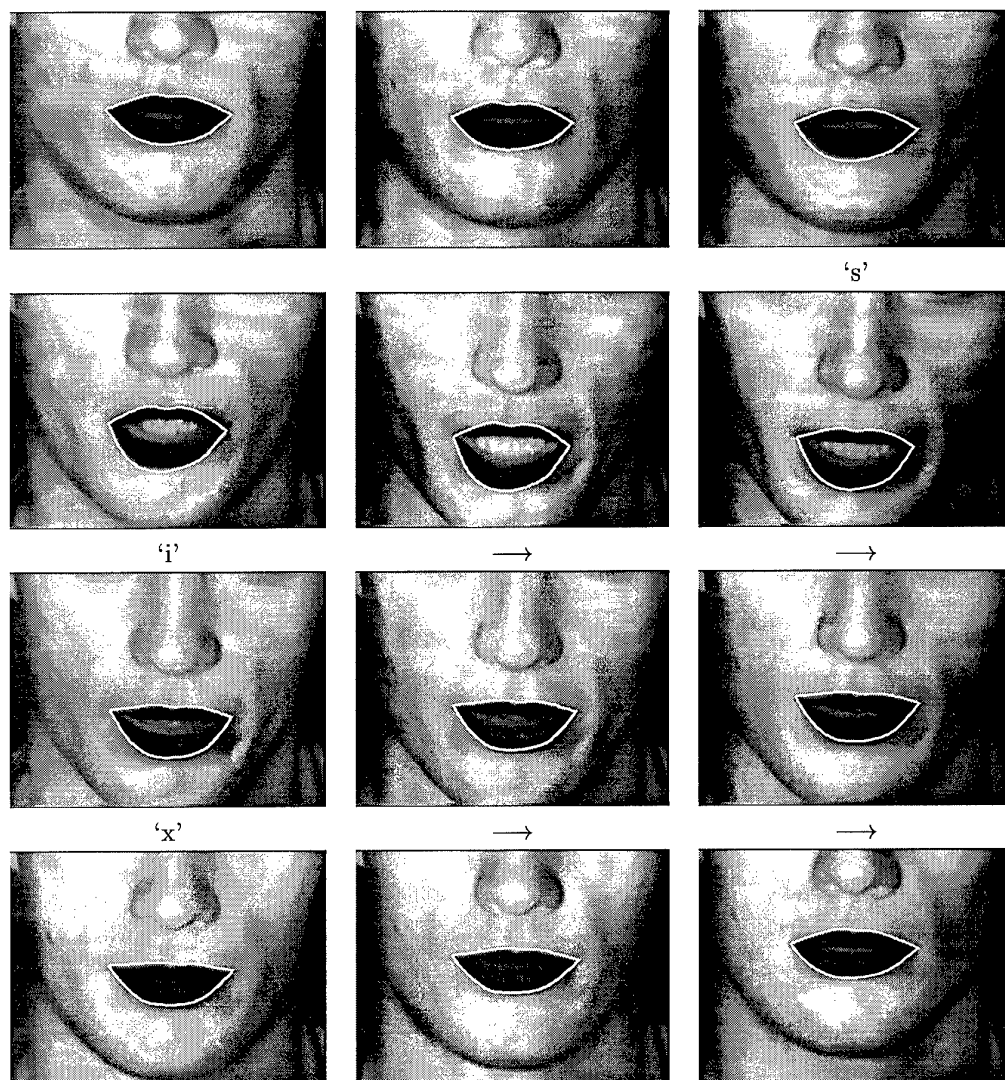
Figure 4.5: *Tracking is robust and accurate even when speech is accompanied by head motion. The speaker nods his head while saying "six". The tracker accurately follows the lips during the nod, including when the mouth opens in concert with the nod as seen in the 'i' (frames 4–6). Note that the nod represents approximately a 45° rotation perpendicular to the image plane (compare frames 6 and 12).*

4.3.1 Feature Detection

The example tracking sequences demonstrate the accuracy achieved using the cosmetically-assisted lip tracker. Qualitatively, one can say that the tracking is very accurate; however, it is also possible to make more quantifiable judgements about its performance. If identical tracking spaces and learnt model dynamics are used, then the quality/accuracy of a tracker is correlated directly with the ability of the feature detection method used to correctly identify and accurately pinpoint features within the image that correspond to the boundary of the tracked object, in this case, the outer lip contour. Specifically, when feature measurements are obtained by searching along one dimensional normals to the contour, the accuracy of the tracker is directly related to the squared spatial measurement error (covariance) of the measurement process.

The one dimensional image measurements at sampled positions s_i, along the contour, where i represents the search line number (section 3.5), are given by z_{s_i}. Suppose the measurement error is normally distributed with zero mean

$$E[z_{s_i} - Hx] = 0 \tag{4.1}$$

and covariance

$$E[(z_{s_i} - Hx)^2] = \sigma_{s_i}^2, \tag{4.2}$$

where H is an observation matrix (3.16) that relates the state x to the measurements z. Then the measurement probability density function, $p(z_{s_i}|x)$, for each of the one dimensional search lines is completely specified by its spatial measurement covariance $\sigma_{s_i}^2$. This measurement covariance can be used as a metric for comparing various feature detection methods, where smaller covariances correspond to more accurate tracking. This metric is closely related to the inverse of the *localisation* criterion described by Canny [28] in his development of an optimal step edge detector.

In general, measurement errors may result from any number of factors including optical and electrical shot noise, the image intensity profile along the normal (foreground texture), background clutter, the intensity variations due to changing illumination and shadows, and the image operator used. In lieu of attempting to model all of the factors contributing to errors in the feature detection process, the measurement covariances for each of the feature detection methods used were determined empirically using simulations on actual, or appropriately modelled, image data.

4.3.2 Edge Detection

The intended effect of the lipstick was to create a dominant edge at the boundary between the outer lip contour and the surrounding skin. This was indeed the case as is evident in figure 4.6 where several intensity profiles for a typical search line are displayed as a function of distance from the predicted contour.
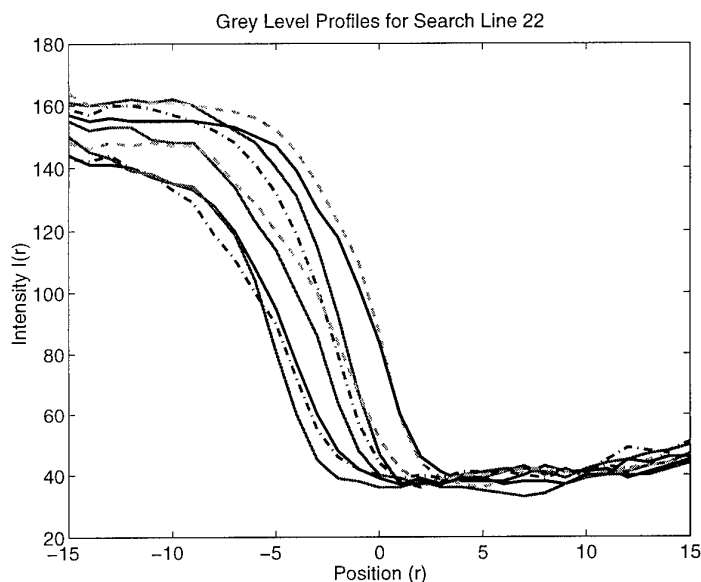

Figure 4.6: *Intensity profiles for a typical search line from the cosmetically-enhanced lips displayed as a function of distance from the predicted contour. Note that application of the lipstick has resulted in a clearly defined edge at the outer lip boundary.*

Accordingly, the image features at the boundary of the cosmetically-enhanced lips are identified by using a Canny edge detector [28] along normals to the predicted contour. As suggested by Canny, the optimal 1D step edge detector was approximated by the first derivative of a Gaussian

$$f(x) = -\frac{x}{\sigma^2} e^{-x^2/\sigma^2} \tag{4.3}$$

which is shown graphically in figure 4.7.

Consistent with standard practice, edges are located by convolving the edge operator, $f(x)$, in this case the first derivative of a Gaussian of width $2W + 1$, with the intensity profile for the given search line, $I(r)$,

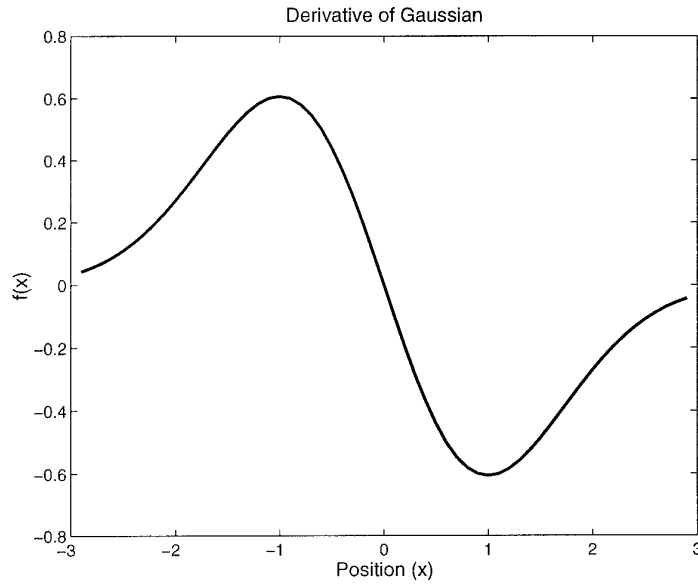$$H(r) = \sum_{x=-W}^{W} I(x - r) f(x). \tag{4.4}$$

Figure 4.7: *The first derivative of a Gaussian with $\sigma = 1$ (the "Canny" edge detector [28]), which was used to locate the lip boundary of the cosmetically-enhanced lips.*

The centre of the edge, z_0, is identified by the maximum of the output of the convolution

$$z_0 = \arg\max_r H(r). \tag{4.5}$$

By modelling the intensity profile $I(r)$ along a given search line as a function $G(r)$, plus zero mean Gaussian noise, η with variance n_0, that is

$$I(r) = G(r) + \eta, \tag{4.6}$$

Canny [28] showed that the covariance of the measurement error, $E[z_0^2]$, for points z_0 near the centre of the edge is given by

$$E[z_0^2] \approx \frac{n_0^2 \displaystyle\sum_{x=-W}^{W} f'(x)}{\left[\displaystyle\sum_{x=-W}^{W} G'(-x)f'(x)\right]^2}. \tag{4.7}$$

However, as Canny was careful to note, $E[z_0^2]$ was derived from the response at only one point (the centre of the edge) and hence failed to take into account the interaction of the responses of the nearby points. A more direct way of computing the measurement covariance σ^2 is to compute it directly using (4.4) and (4.5).

Typically, edge profiles are modelled as step edges, although it is evident from figure 4.6 that these edges look more like error functions [82], where

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt. \tag{4.8}$$

Accordingly, the edge profiles for the lip boundary were modelled as scaled and shifted erf functions

$$G(x) = G_0 - A \, \text{erf}(x/\sigma_e). \tag{4.9}$$

4.3.3 Feature Measurement Error

After fitting the search line (figure 4.6) to the model (4.9), the measurement process was simulated using equations (4.4) and (4.5). The result of one simulation is shown in figure 4.8. It is seen that the image noise has resulted in the marked edge centre ($r = 1$) being one pixel in error from the true edge centre ($r = 0$), although the sharply peaked convolution sum (bottom figure) illustrates the good localisation of the Canny edge detector. The accuracy of this detector is verified by simulating the edge detection measurement process one thousand times. The resultant mean and covariance of the measurement error were found to be

$$\mu = -0.02 \text{ pixels} \quad \text{and} \quad \sigma^2 = 0.3270 \text{ pixels}^2 \tag{4.10}$$

which is consistent with the assumed zero mean Gaussian. This extremely low measurement error (less than one pixel^2) confirms the qualitative observation that the cosmetically-assisted lip tracker was extremely accurate and robust.

4.3.4 Summary

This tracker demonstrates that the dynamic contour framework is well suited for developing shape and motion models which permit accurate, real-time lip tracking. Furthermore, the tracker is agile enough to follow the complex deformations of rapidly moving lips, and later audio-visual recognition experiments (chapter 6) confirm that the outline of the lips is a rich source of information. The high accuracy of this tracker also permitted detailed analysis of the lip motions present in natural speech which is discussed in section 6.1.

While this tracker proved to be extremely useful for developing shape and motion models, it would be undesirable to require users to wear lipstick in a commercial speech recognition setting. However, one can think of many instances where this limitation would not be unreasonable, such as, actor-driven facial animation [6] and using trained speakers for the deaf.
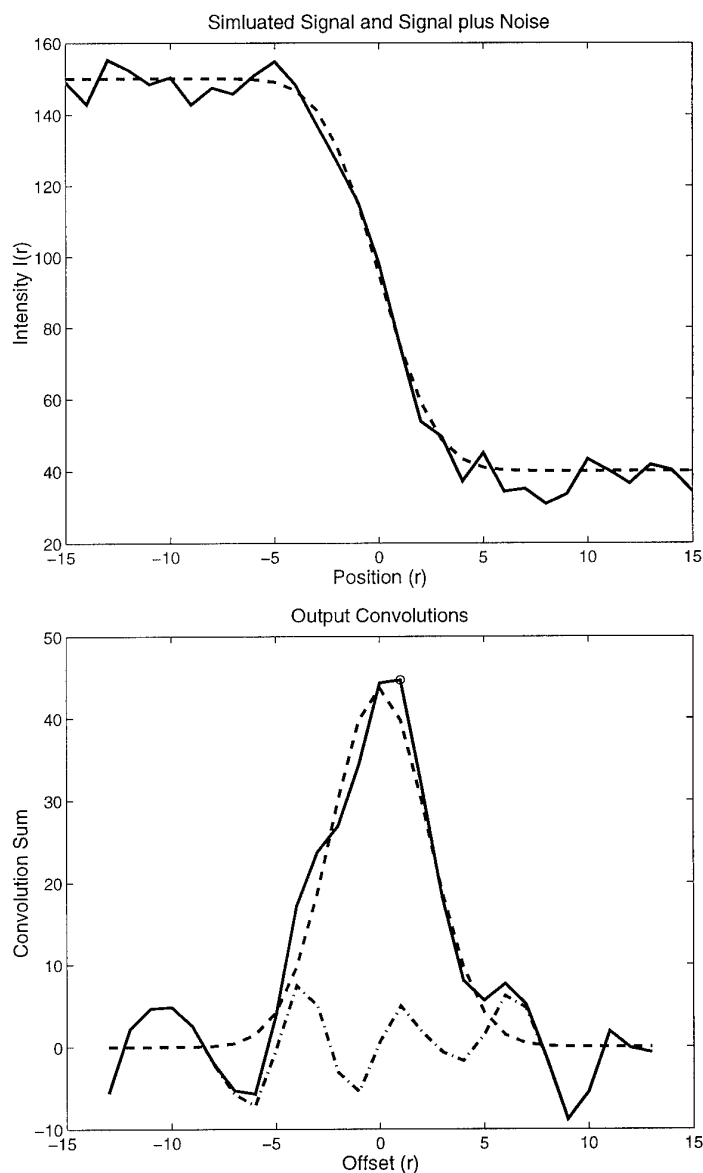
Figure 4.8: *Simulation of the edge detection measurement process. The top graph shows the noise free edge (dashed line) along with the noisy edge (solid line). The bottom graph is the result of the edge detection convolution where the dashed line is the noise free edge, the dash-dot line is the output of the convolution with the Gaussian noise only, and the solid line is the simulated noisy edge. The 'O' marks the location of the identified edge ($r = 1$). Note that the image noise has caused the output peak to be smoother, and that the marked edge centre is one pixel in error.*

Having established that real-time tracking of the lips from the frontal view was indeed possible within the dynamic contour framework, additional trackers were developed which used advanced feature detection mechanisms to address the lack of image features around

the lip contour in unadorned lips.

4.4 Correlation Matching

In order to identify the lip boundary in poorly contrasted facial images (figure 4.1) alternative approaches to edge detection were investigated. The first method examined was correlation matching. Correlation matching is a widely used feature detection scheme [3, 90, 7], because of its simplicity and its similarity to the matched filter from communication theory [35]. It is rooted in straightforward Euclidean distance template matching.

For a 1D line segment, the squared Euclidean distance between a search line $I(x)$ and a reference template $T(x)$ is

$$D(r) = \sum_{x=-W}^{W} (I(x+r) - T(x))^2 \tag{4.11}$$

which when expanded gives

$$D(r) = \sum_{x=-W}^{W} I^2(x+r) - 2I(x+r)T(x) + T^2(x). \tag{4.12}$$

Closer inspection of (4.12) reveals that the $T^2(x)$ term is independent of the position r along the line segment and that $\sum I^2(x+r)$ is merely the image energy in the window about the region $I(r)$. If the variation in image energy can be assumed to be small as a function of position r, then it too can be assumed to be a constant and hence ignored in the total distance calculations. Thus the total squared distance, $D(r)$, is a minimum when $\sum I(x+r)T(x)$ is a maximum. Accordingly, finding the minimum squared Euclidean distance is equivalent to finding the maximum cross-correlation $C(r)$, where

$$C(r) = \sum_{x=-W}^{W} I(x+r)T(x). \tag{4.13}$$

However, when the correlation region is small, which is often the case in the dynamic contour lip trackers, the assumption that the image energy is independent of position is invalid. To compensate for this, normalised intensities can be used by dividing (4.11) by the image energy

$$\sum_{x=-W}^{W} I^2(x+r).$$

There are, however, drawbacks when using normalised intensities; specifically, there is a loss of information in that the raw intensity values are no longer available. This can, and

often does in practice, lead to false matches within the image that do not correspond to the lip boundary. In addition, even when the normalisation process results in only minor degradation of the pertinent information, the use of finite length sequences can result in maxima of the correlation sum that do not correspond to the best alignment of the template with the image data. This phenomenon is evident in figure 4.9.
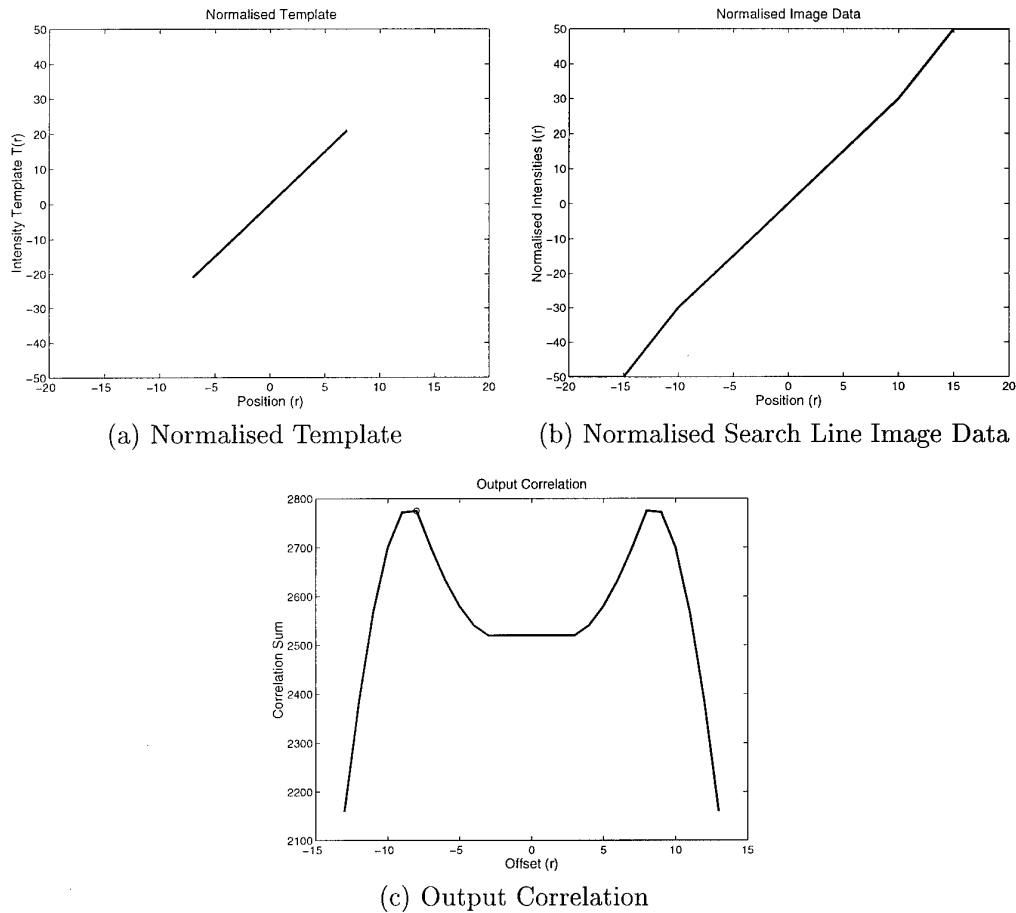

(a) Normalised Template   (b) Normalised Search Line Image Data

(c) Output Correlation

Figure 4.9: *The maximum of the normalised cross correlation does not coincide with the best alignment. The normalised template (a) and the image data (b) are identical in the range $-8 \leq r \leq 8$. Thus, the best alignment (maximum of the correlation sum (c)) should occur at $r = 0$; however, because finite length sequences are used, the maximum of the output correlation (denoted by 'O', $r = -8$) does not correspond to this best alignment.*

Another limitation of correlation matching is its inability to incorporate known variations in intensities resulting from the tracking process into the reference templates. Due to these limitations of correlation matching, further study on it was discontinued in favour of a feature detection approach based on a more general statistical analysis of the intensity

profiles using a weighted Euclidean (or Mahalanobis [46]) distance measure.

4.5 Tracking using Statistical Modelling

A more general approach to the correlation matching technique discussed in the previous section is to use a data-driven approach. That is, the information content of the region surrounding the contour boundary can be captured using statistical models of the grey-level appearance along each of the search line normals. During tracking, the statistical templates, $\mathbf{T}$, can be compared to the image profiles, $\mathbf{I}(r)$, using a squared Mahalanobis distance metric, $M(r)$, which weights the distance between the template profiles and the image data by their uncertainty

$$M(r) = (\mathbf{I}(r) - \mathbf{T})^T P^{-1}(\mathbf{I}(r) - \mathbf{T}). \tag{4.14}$$

Here, $\mathbf{T}$ is the template, an L-vector of pixel intensities, and r is the signed distance from the lip boundary along the outward pointing normal. $\mathbf{I}(r)$ is an L-vector of pixel intensities along the normal, centred at r. It can be shown [46] that $M(r)$ is proportional, up to an additive constant, to the log-probability of a multi-variate normal distribution centred at $\mathbf{T}$ with covariance P, and hence that the minimum Mahalanobis distance represents the maximum likelihood location of the image feature. Such feature search techniques have been used previously in non-real-time applications by Cootes et al. [34, 83] and by Rowe [121, 122]. However, since the computational burden associated with the learning of the statistical models is accomplished off-line, they can be effectively used to locate image features within a real-time framework.

4.5.1 Learning the Statistical Templates

The grey-level intensities along a search line profile in a particular region about the mouth, in general, will be a function of the illumination and the texture of the underlying lips and skin. However, other sources of variations will also be present due to shadowing, image noise, and imperfections of the B-spline fit to the lip contour. Statistical models can be used to represent and capture all of these uncertainties and variations. These templates and their corresponding uncertainties (P in 4.14) can in turn be incorporated into the distance measure. A separate intensity template, $\mathbf{T}_i$, can be created for each search line at sampled positions, s_i, along the contour (see figure 4.10). This data-driven approach allows the salient aspects particular to each intensity profile to be captured by the statistical models.

The search line templates are learned from training image sequences using a bootstrap procedure. An initial set of intensity profiles $\mathbf{I}_{i,t}$ is obtained by hand-fitting splines to the
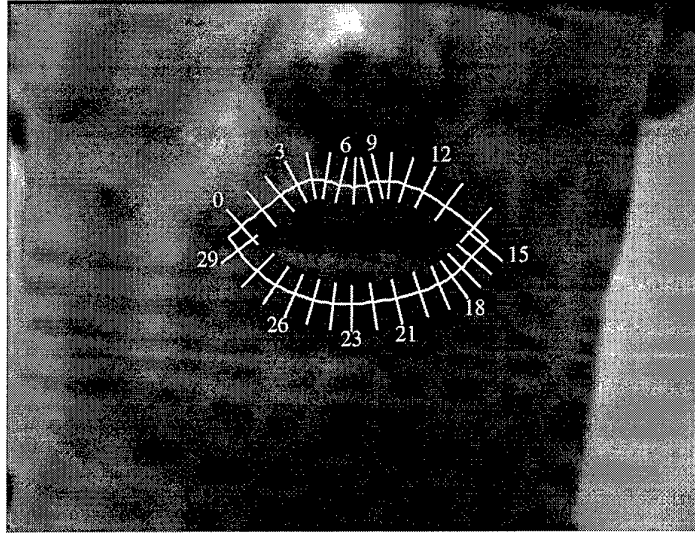
Figure 4.10: *A separate statistical intensity template is created at each of the sampled positions along the lip contour. These search lines are labelled clockwise from 0 − 29. Both the mean intensities and the covariances are learnt from training image sequences.*

outer lip contour in several images and then extracting the grey-level intensities centred at the contour. Initial estimates for the template means and covariances are computed using an unbiased estimator,

$$\mathbf{T}_i \;=\; \frac{1}{N}\sum_{t=1}^{N}\mathbf{I}_{i,t} \tag{4.15}$$

$$P_i \;=\; \frac{1}{N-1}\sum_{t=1}^{N}(\mathbf{I}_{i,t}-\mathbf{T}_i)(\mathbf{I}_{i,t}-\mathbf{T}_i)^{T}. \tag{4.16}$$

These templates are used in the measurement process via (4.14) in a simple tracker to locate and identify image features along the search lines. This permits the acquisition of additional intensities profiles which can be obtained from tracking conditions representative of those likely to be encountered in future tracking sessions. These new profiles are then used to update the means and covariances accordingly.

A point to be made is that the use of a full covariance matrix P requires the estimation of $\mathcal{O}(L^2)$ parameters and, more importantly from a real-time tracking perspective, $\mathcal{O}(L^2)$ multiplications and additions (4.14) for each distance calculation. A simpler model can be utilised if the pixels along the search line are assumed to be statistically independent, which results in a diagonal P and reduces both the number of parameters to be estimated and the number of required arithmetic operations to $\mathcal{O}(L)$. The results of the template learning process, using this simplifying assumption, for a prototypical search line are shown
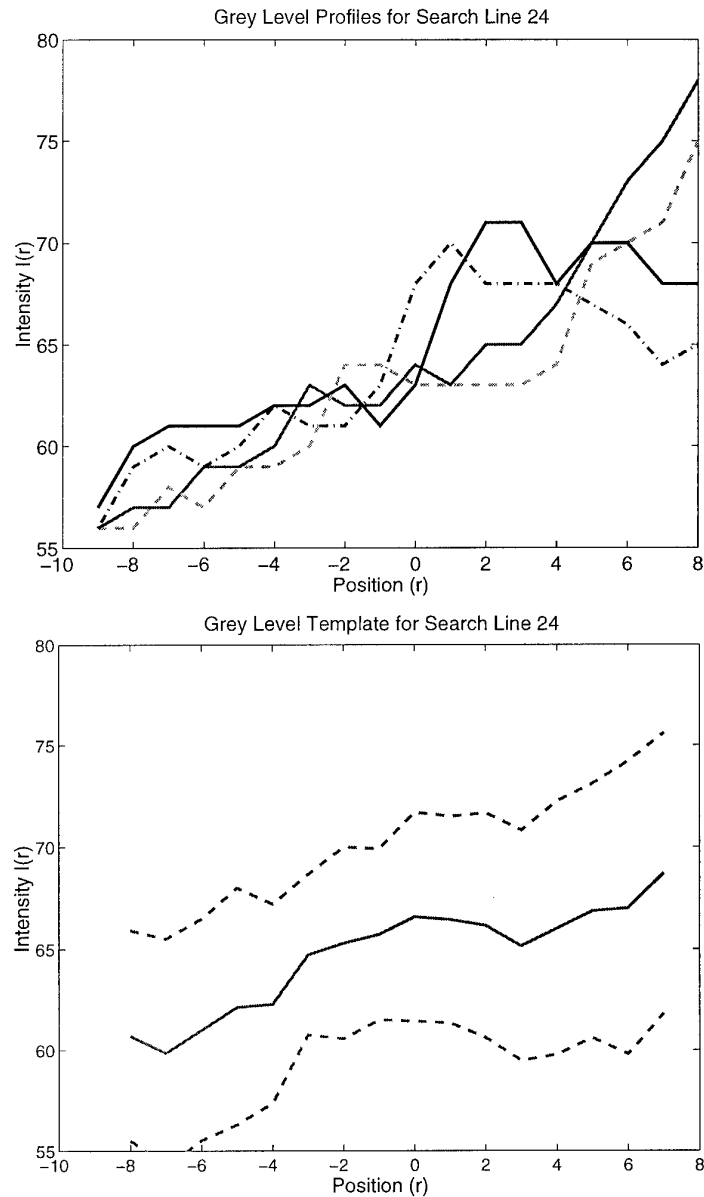
in figure 4.11.

Figure 4.11: *Prototypical grey-level intensity profiles along an image search line and the learnt statistical template plotted as a function of distance from the lip contour. The solid line represents the mean intensity and the dashed lines ±1 standard deviation change in intensity. The absence of a sharp jump or sudden change in the intensity profiles also clearly demonstrates why standard edge detection methods are inappropriate for identifying the lip boundary which corresponds to $r = 0$ and is not distinguished by an appreciable edge.*

In this work, only 1D intensity profiles were used, although this framework can be easily

extended to regions of arbitrary shape near the sampled contour positions. Similarly, it is not necessary that raw intensities be used; experiments were done using normalised intensities, as well as the gradient of the intensities, and the normalised gradient. Some success was achieved using normalised intensities, although the loss of information due to normalisation led to excessive false matches, degrading the overall performance of the tracker. Somewhat less success was achieved using gradient and normalised gradient templates, because in regions with relatively constant gradient, differentiating the intensity profile resulted in the loss of the salient information. Such regions exist along the lower lip which is illustrated in figure 4.12.
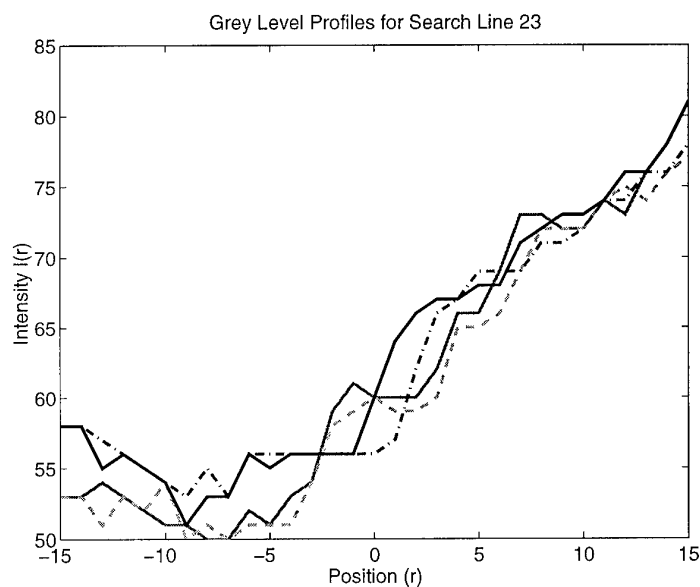
Figure 4.12: *Representative intensity profiles along the lower lip. Since the profiles have relatively constant gradient, differentiating them results in the loss of the salient information, which degrades tracking performance when they are used in the feature search routines.*

Rowe et al. [121, 122] have recently proposed a more general statistical modelling technique than the one described here. They make no assumptions about the form of the distribution of the grey-level intensities along the template (other than statistical independence between adjoining pixels) and instead learn the intensity probability distributions directly from the image data. They also permit non-linear warps between the intensity templates and the image search line data, as opposed to only translational shifts. However, their added generality results in an enormous increase in computational complexity, precluding their use in real-time tracking applications. The modest assumptions made here concerning

the nature and form of the data strike a more appropriate balance between generality and computational complexity.

4.5.2 Feature Measurement Error

As discussed in section 4.3.1, the accuracy of a tracker is directly related to the quality of the feature detection method used. More specifically, the degree to which the feature detection method results in a unimodal, sharply peaked, measurement probability density function determines the tracker's ability to correctly identify and precisely locate the lip boundary.

In section 4.3.3 it was shown how to determine the measurement probability density function $p(z_{s_i}|x)$, where x is the true position of the image feature and z_{s_i} represents the observed location, ie. the *detected* feature. A similar analysis is given here where the measurement densities, $p(z_{s_i}|x)$, are obtained empirically by simulating the Mahalanobis feature detection process.

The measured 1D position, z, of a feature is given by the minimum squared Mahalanobis distance

$$z = \arg \min_r M(r), \tag{4.17}$$

where $M(r)$ is given by (4.14).

By modelling the variations in the grey-level intensities, $\mathbf{I}$, in the surrounding region of the template profiles, $\mathbf{T}$, the measurement densities can be estimated directly from equations 4.14 and 4.17. The statistical variations, P_B, in the surrounding regions can be learned in the same manner as was used for learning the covariances of the intensity template (section 4.5.1). Artificial image data can then be generated using these learnt distributions

$$\mathbf{I}(r) = \bar{\mathbf{I}}(r) + \eta \tag{4.18}$$

where $\bar{\mathbf{I}}$ are the mean profile intensities and η is a normally distributed L-vector with zero mean, and covariance P_B. The measurement densities can then be estimated from equations 4.14 and 4.17 using this simulated data. One such simulation for a sample search line is shown in figure 4.13.

The most prominent observation from these graphs is that the output squared Mahalanobis distance (graph (c)) is flat over a fairly large region, indicating that errors in localising the position of the image feature are likely. This is in contrast to the sharply peaked output of the cosmetically-assisted lip tracker (figure 4.8) where the lipstick resulted in dominant edge features that were robust to image noise.
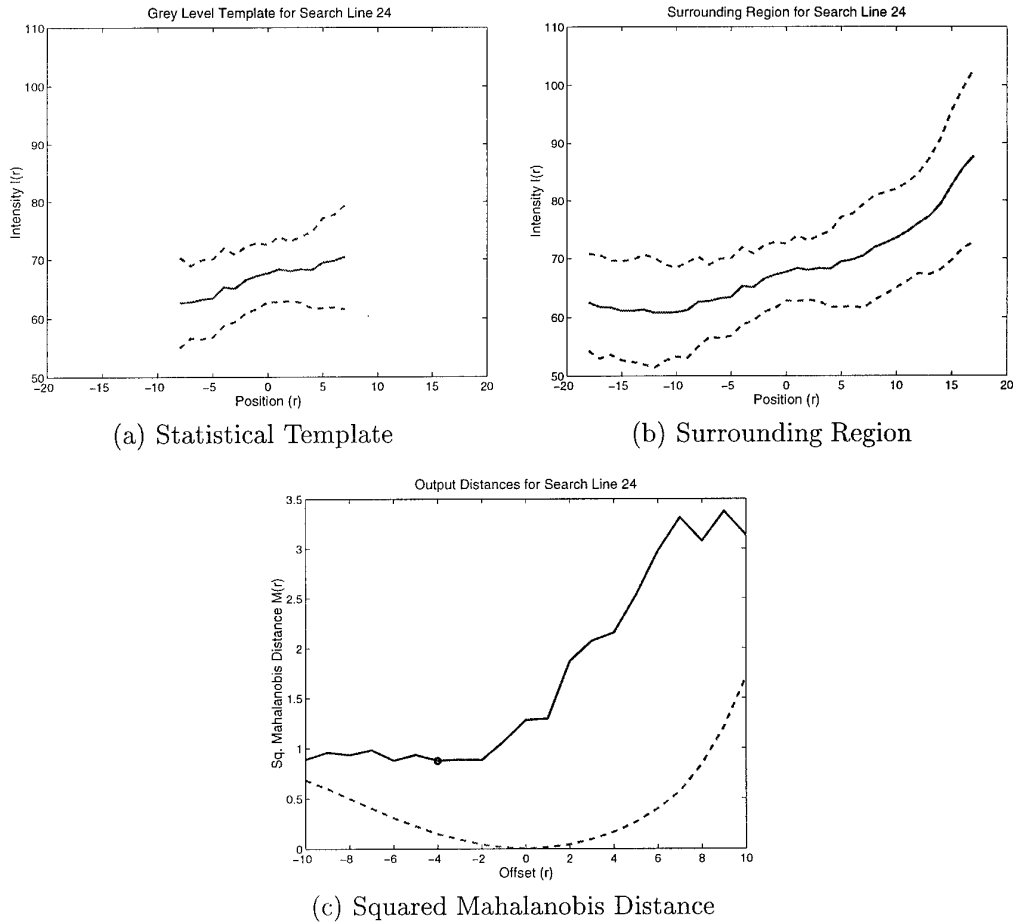
(a) Statistical Template

(b) Surrounding Region

(c) Squared Mahalanobis Distance

Figure 4.13: *Simulation of the Mahalanobis feature detection process. The first graph (a) shows the statistical template along with ±1 standard deviation from the mean intensities, where $r = 0$ denotes the position of the contour. The second graph (b) shows the search line image data in the vicinity (±20 pixels) of the lip contour along with the variations in the neighbouring pixels (dashed line). The last graph (c) shows the squared Mahalanobis distance between the template and the noise-free image search line (dashed line) and between the template and the simulated noise corrupted image profile (solid line). Note that for this search line the output distances (dotted line) are relatively flat for $-10 \leq r \leq -2$ demonstrating its susceptibility to image noise. Contrast this curve to the sharply peaked output of the cosmetically-assisted lip tracker (figure 4.8). Addition of the noise (solid line) results in misidentification of the lip boundary (marked as 'O', $r = -4$) which is 4 pixels in error.*
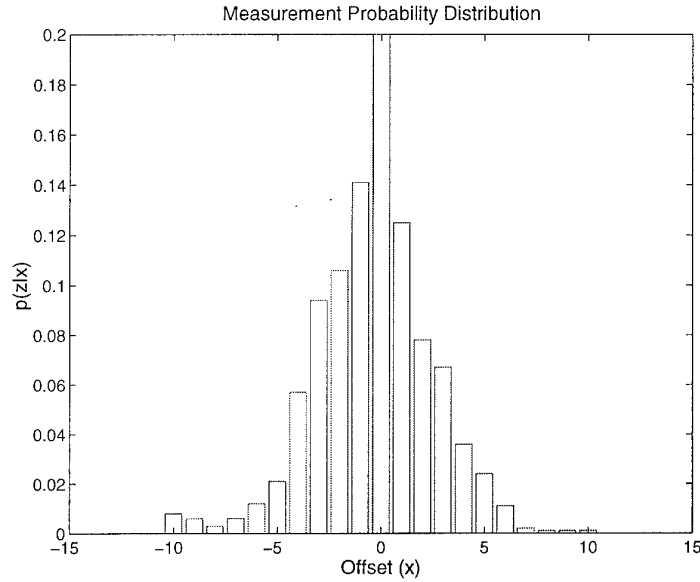
Figure 4.14: *The measurement probability distribution from simulation of the Mahalanobis measurement process. The distribution is roughly Gaussian with mean $\mu = -0.387$ pixels and covariance $\sigma^2 = 7.865$ pixels^2.*

The measurement probability distribution, $p(z|x)$, for the search line shown in figure 4.13 was determined empirically by simulating the feature detection measurement process one thousand times. The resultant measurement density is shown in figure 4.14 as a function of offset from the true position. The measurement distribution is roughly Gaussian with mean and covariance

$$\mu = -0.387 \text{ pixels} \quad \text{and} \quad \sigma^2 = 7.865 \text{ pixels}^2.$$

Not surprisingly, the covariance of the measurement noise is much higher than the sub-pixel accuracy of the cosmetically-assisted lip tracker, where $\sigma^2 = 0.327$ pixels^2 (4.10). When compared to this sub-pixel accuracy, the 7.865 pixels^2 measurement covariance of this detector may appear large; however, when translated into image coordinates, the measurement error is relatively small. This is illustrated in figure 4.15 where some of the detected features around the lips are several pixels in error from the true lip outline, although the resultant least-squares fitted spline approximates the lip contour well.

4.5.3   Tracking

The statistical templates can now be used as feature detectors in the dynamic contour tracking framework. A short excerpt from a tracked sequence of the word "seven" (figure 4.16)
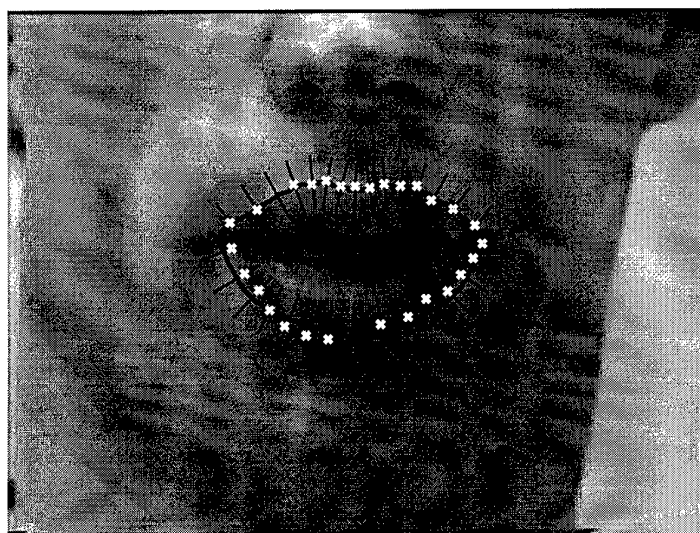
Figure 4.15: *Using statistical template matching results in the detection of features along the lower lip, despite the poorly contrasted lip-skin boundary. As discussed in the text, the measurement error of the feature detector results in several of the detected features being displaced from the actual lip outline, especially along the lower lip. However, many of the features have been precisely identified, and the resultant least-squares fit approximates the lip outline well. For reference: each of the search lines (thin black lines) are 25 pixels in length.*

demonstrates the tracking accuracy achieved using the statistical models in conjunction with the Mahalanobis distance measure.

The resultant accuracy of the tracker suggests that the statistical models have captured attributes of the intensity profiles necessary for identifying the lip boundary. In addition, a strength of the modelling approach is that variations due to lighting changes can be directly incorporated into the statistical templates by training on intensity profiles obtained from images under various illumination conditions. In applications where there are not gross changes in lighting, such as speech recognition in an office environment with fluorescent lighting, the statistical detectors adequately handle this variability. This is illustrated in figure 4.17.

Although the statistical template matching feature detection method has been presented in reference to lip tracking, there are many other tracking domains where this modelling approach could be used effectively. For instance, the tracking of the human heart in ultrasound imagery is receiving increased attention as a method for detecting heart abnormalities [72]. However, there is poor contrast along the boundary of the heart due to the low signal-to-noise ratio of the ultrasound images. The employment of statistical feature detectors should adequately capture the salient information of the heart boundary.
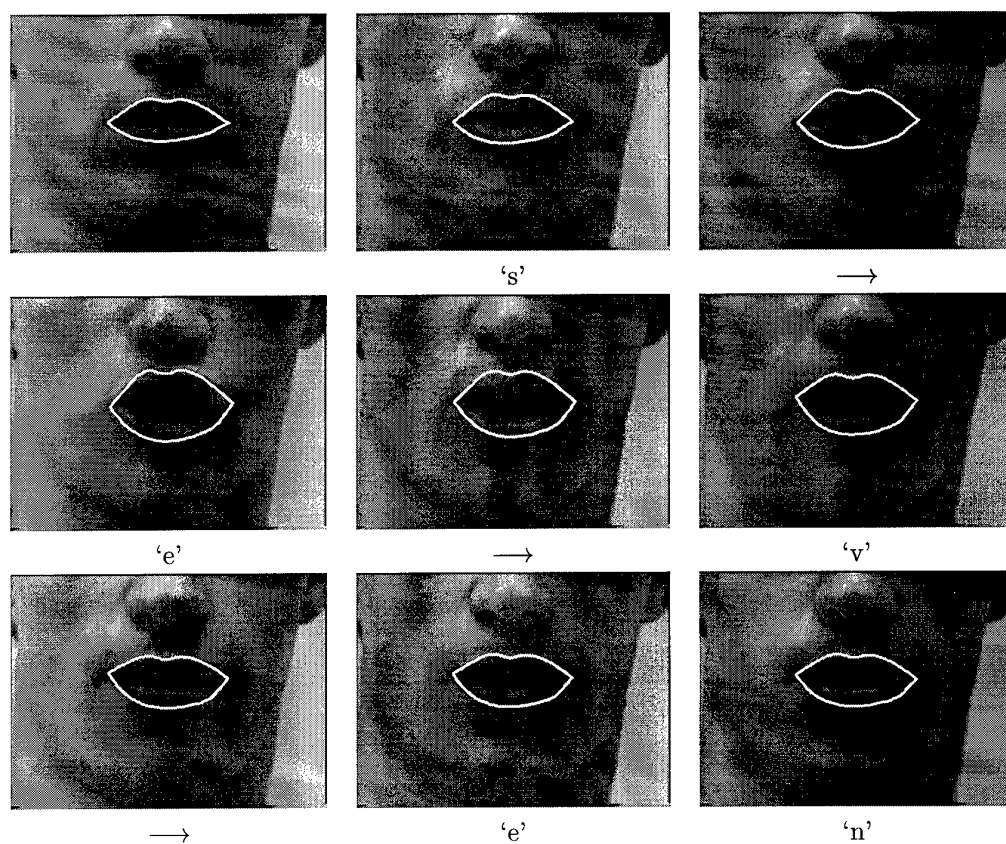
Figure 4.16: *Tracking the word "seven" using the statistical modelling feature detection method. Snapshots taken approximately every 60 ms. The white line represents the position of the contour after the measurements have been assimilated. The tracker accurately follows the lips throughout the entire sequence; however, in frames 3 and 4, it can be seen that the contour has not fully deformed to the curled upper lip as the speaker is transitioning from 's' to 'e'.*
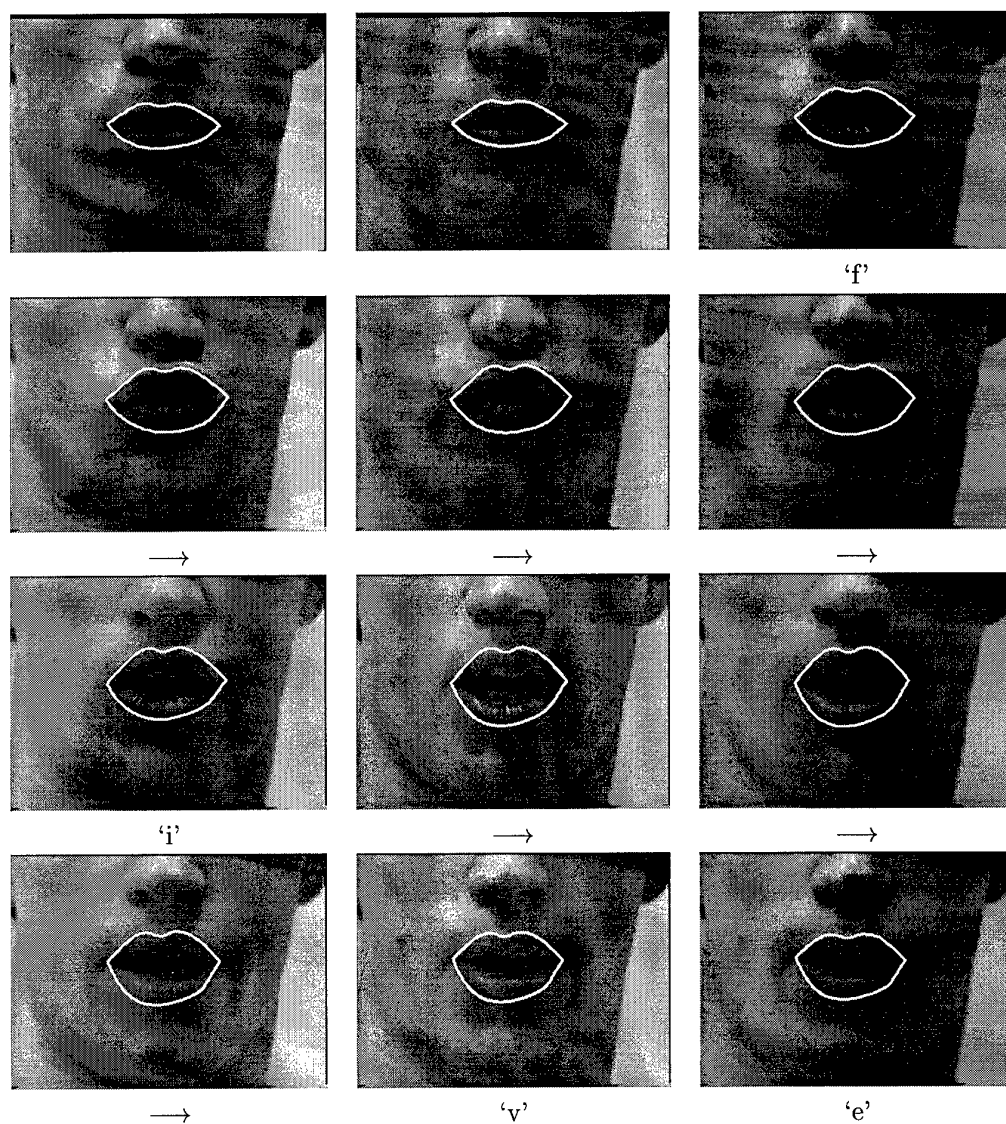
Figure 4.17: *Tracking is robust to small lighting changes as typically encountered in an office environment. Snapshots taken approximately every 80 ms. "Five" is accurately tracked throughout the sequence, although there are slight misalignments along the upper lip in frames 3–5.*

Another application is the tracking of objects in natural scenes against varying backgrounds. The grey-level appearance of the tracked object could be represented using statistical templates with relatively small variances, so-called foreground modelling, while the portion of the template representing the background clutter would have large variances indicating the uncertainty associated with the background. During tracking, foreground features should reliably match the foreground portion of the profile templates, while the background clutter is absorbed by large variances in the background portion of the templates.

4.6 Conclusion

The accurate lip tracking performance attained in this chapter demonstrates the power of the real-time dynamic contour tracking framework. The difficult task of tracking rapidly deforming, articulating lips was tackled by first developing appropriate shape and motion models which were *learnt* from continuous speech sequences. Next, methods for identifying the lip boundary in grey-level images were investigated. It was found that the distracting effects of the teeth, in conjunction with the difficulty of accurately locating the corner of the mouth, made tracking the outer lip contour preferable to tracking the inner one. Correlation matching was shown to be ill-suited for identifying the weakly contrasted lip boundary. It was then demonstrated that statistical models could successfully capture the salient information of the intensity profiles along the search lines of the lip contour. This enabled reliable identification of the lip-skin boundary. When the statistical feature detectors were used in conjunction with the shape and motion models, the result was accurate, real-time tracking of unadorned lips.

5

Audio-Visual Recognition Systems

This chapter provides an overview of the various sub-systems comprising two audio-visual speech recognisers. These recognisers were developed in order to assess the ability of visual information extracted from the lip contour to provide robust speech recognition. The recognition experiments conducted using these recognisers are presented in the next chapter.

The first recogniser uses a dynamic time warping (DTW) matching algorithm to account for the non-linear temporal variations inherent in speech, while the second uses stochastic modelling in the form of continuous density Hidden Markov Models (HMMs). Both systems utilise the dynamic contour trackers presented in the previous chapter. The DTW system was designed for isolated-word recognition tasks only, while the HMM system was implemented as a continuous-word recogniser in order to permit investigation into some of the practical problems facing commercial audio-visual speech recognisers.

Much of the pre-processing and feature extraction steps are identical for the two systems. As such, an explanation of the sub-systems common to both recognisers is given only in the Dynamic Time Warping section.

5.1 Dynamic Time Warping Recogniser

Dynamic Time Warping is a dynamic programming technique that finds the minimum distance between two sequences given a local distance measure and global path weightings [111]. Using local optimisation, DTW finds the optimal warping that results in the

minimum cumulative distance between the two sequences. All possible paths, subject to the global constraints, are considered. When matching different repetitions of the same recognition unit (phoneme, syllable, or word), DTW should compensate exactly for the speaking rate variations. There is no such correspondence when comparing different words — the DTW algorithm merely finds the shortest distance between the sequences. However, in theory, the distance between dissimilar words should be greater than that between like ones.

The DTW recogniser consists of two parts. The first is the "training" system where reference templates are created from known and labelled utterances and the second is the "recognition" or evaluation system where unknown words are processed and then matched against the stored reference templates. Block diagrams of the training and recognition sub-systems are shown in figures 5.1 and 5.2, respectively. The subsequent sections detail the different portions of the DTW system.

5.1.1 Segmentation

The first task in any isolated-word recognition problem is identification of the word boundaries within the sample utterance, that is, endpoint detection or word segmentation. In situations where the word is spoken in isolation in a quiet environment, voice activated detection (VAD) methods seem to work well [67]. However, even in quiet environments, the problem is more difficult than might first appear, for it is common for speakers to precede (or follow) words with lip smacks and/or spurious noises which make accurate endpoint detection difficult. Some words (such as eight and contemplate) have periods of silence within them that can be easily mistaken for word boundaries. Furthermore, VAD schemes are often unreliable in endpointing words beginning or ending with weak sounds, such as /f/, /s/, and /k/, where the level of the noise may be greater than that of such sounds. Complicating matters further, speakers occasionally omit the stop consonant burst on words ending in unvoiced stops (/p/, /t/, and /k/) [67].

In this work, since the acoustic signal was recorded in noise free conditions (artificial noise was added later), segmentation was accomplished using only the acoustic channel (see figure 5.1). The endpoint detector calculated the average power in a sliding window and compared it to a pre-set threshold. The beginning (ending) of the utterance was identified as the point where the average power first exceeded (dropped below) the threshold. To ensure that no leading or trailing sounds were omitted by the detector, the endpoints were extended by 100 ms (800 samples), forward and backward. The identified endpoints were then used to segment the visual features to ensure good synchronisation between the two
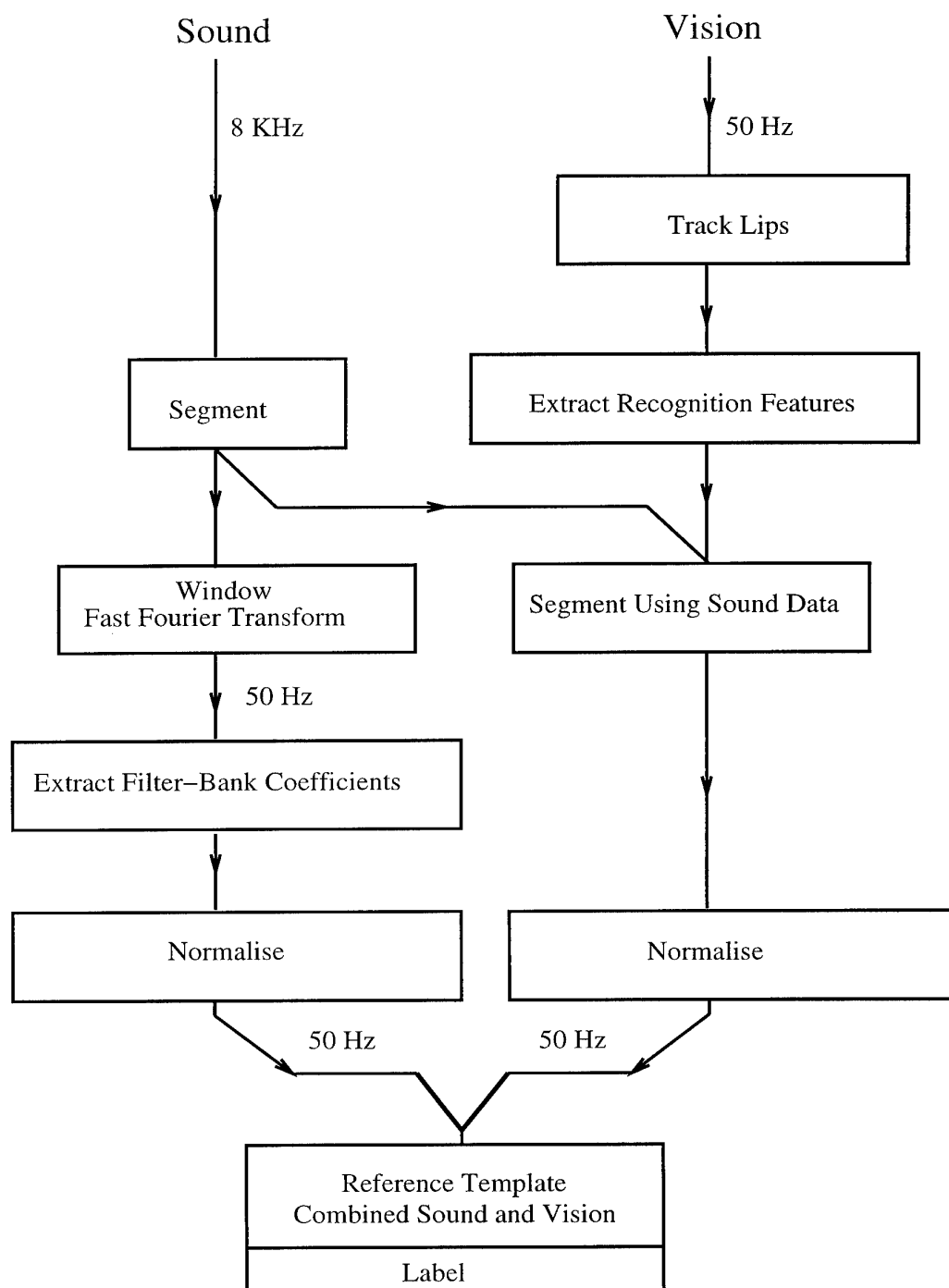
Sound   Vision

8 KHz   50 Hz

Track Lips

Segment   Extract Recognition Features

Window
Fast Fourier Transform   Segment Using Sound Data

50 Hz

Extract Filter–Bank Coefficients

Normalise   Normalise

50 Hz   50 Hz

Reference Template
Combined Sound and Vision

Label

Figure 5.1: *Block diagram of the DTW training sub-system. The raw audio and visual data are processed and relevant recognition features are extracted and combined into reference templates which are then used for pattern matching.*

Component Weightings

$$(W_1 \quad W_2) \quad W_3) \quad \bullet \ \bullet \ \bullet \quad (W_n)$$

Reference Templates $\longrightarrow$ Pattern Matching Dynamic Time Warping $\longrightarrow$ Recognised Word

word 1a
label

word 1b
label

word 2a
label

word 2b
label

Processed
Unknown Word

•
•
•
•
•

word Wa
label

word Wb
label
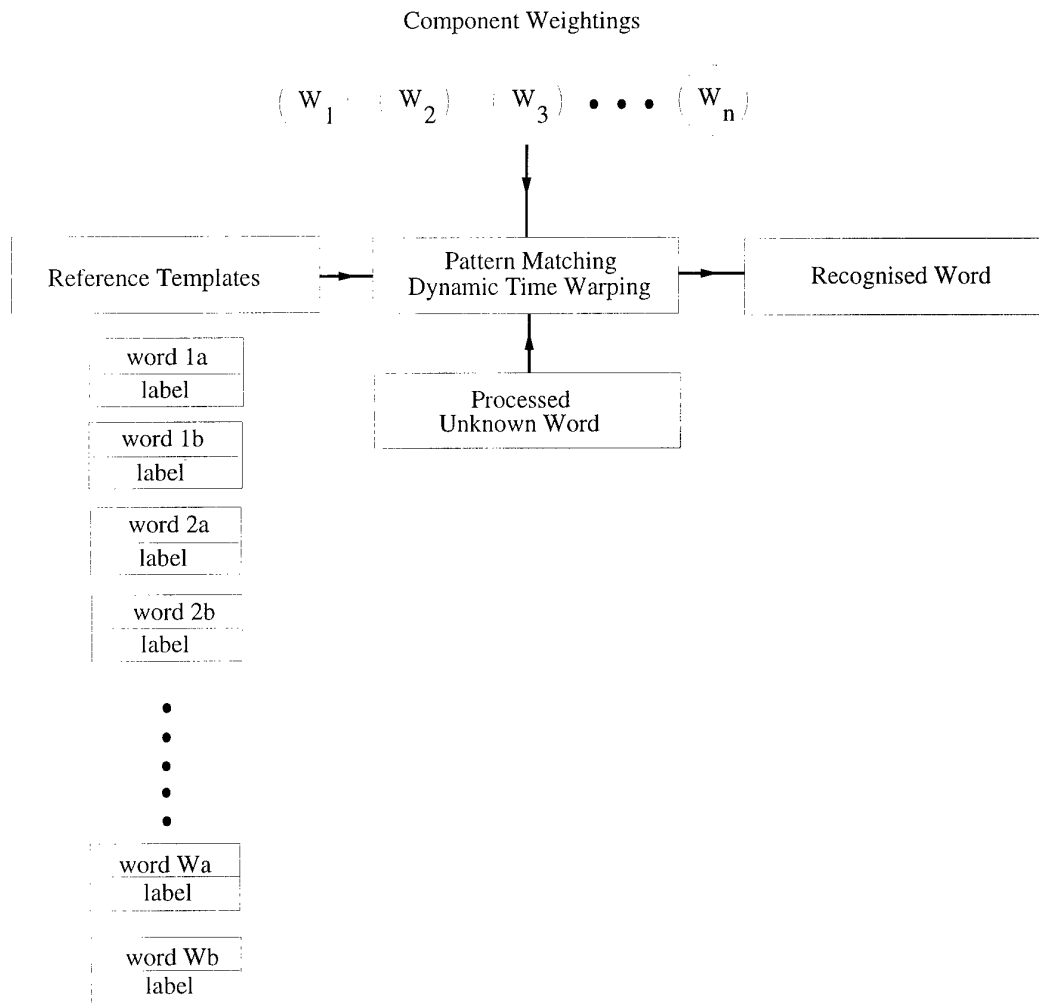
Figure 5.2: *Block diagram of the DTW recognition sub-system. Unknown words are processed and then compared against stored reference templates using the DTW algorithm. The closest match is identified as the "recognised" word. Component weightings can be varied to give greater significance to different features. For example, experiments were conducted where the weightings were varied between the sound and vision components to compensate for the reduced information content of noise-degraded sound.*

modalities. This method of segmentation also ensured the capture of any pre-positioning of the visual articulators prior to speech generation.

Ideally, information from both the audio and visual channels should be used to aid in identification of word boundaries. Indeed, as it is known that even small errors in endpoint detection (≈ 60 ms) result in degraded recognition of isolated digits [111], word segmentation represents a natural domain for the fusion of audio-visual data, particularly in noisy environments. Surprisingly, this area has received only limited attention [89]. This is most likely due to the difficulty of separating and decoupling global head motion from local lip motion and of distinguishing non-speech facial movements, like smiling, from those due to speech. However, acoustic cues should certainly help solve the latter problem, and recent work in the decoupling of pose and expression [78, 6] holds promise for the former.

5.1.2   Audio Feature Extraction

An essential part of any recognition system is the extraction of features that reliably represent the objects in the data set. The features must compactly represent the data in a form suitable for recognition. For speech signals, the features typically result from spectral processing of the acoustic waveform [113, 111].

Consistent with standard practice in computer speech recognition, the speech signal was assumed to be piecewise stationary; that is, it is assumed that over a short time interval the spectral characteristics of the speech signal do not change. This allows the speech waveform to be broken into short segments (called *analysis windows*) which can be analysed independently. Typically, overlapping windows are used, resulting in a new feature vector every *frame interval*. It is important that the windowing interval be sufficiently short (10–40 ms) that only minor shape variations occur in the vocal tract.

In order to minimise signal discontinuities at the beginning and ending portion of each frame, the speech signal was multiplied by a Hamming window, $h(n)$, of length $N = 256$ (32 ms) where

$$h(n) = 0.54 - 0.46 \cos \left(\frac{2\pi n}{N-1} \right), 0 \leq n < N. \tag{5.1}$$

The Hamming window provides adequate resolution in the frequency domain as well as sufficient attenuation in the side lobes. A 37.5% overlap of the windows was used (96 time samples) resulting in a 20 ms frame interval. This frame rate was chosen to coincide with the 50 Hz video rate to facilitate integration of the two modalities without additional sub-sampling or linear interpolation.

The processed speech signal was then passed through a bank of bandpass filters. The

centre frequencies and bandwidth for each of the filters was determined using the mel scale, which results in bands containing equivalent amounts of linguistic information [111]. The spacing of the mel scale is based on perceptual studies and is given by

$$\text{Mel}(f) = 1127 \ln \left(1 + \frac{f}{700}\right),$$

which encapsulates the reality that humans are better at distinguishing sounds at lower frequencies than at higher ones. A measure of the spectral energy in each of the frequency bands was obtained by computing the fast fourier transform (FFT) on the windowed signal. The resultant 128 frequency sample points were then partitioned into 8 bands, leaving 8 mel-scale filter-bank coefficients occurring at a rate of 50 Hz (one feature vector per frame).

5.1.3 Visual Feature Extraction

There is currently no consensus among the speechreading community as to which features most efficiently capture the linguistically informative attributes of the visual signal. It is generally accepted that the lips, teeth, and tongue all contain linguistically relevant information, with the lips being the most informative [25, 131, 84]. Some researchers have demonstrated that the lips alone can carry up to two-thirds of the speech intelligibility conveyed in images of the face [94, 9]. Typically, the lip attributes deemed most informative are the height and width of the oral cavity, and the degree of lip rounding (puckering) [99, 48, 55, 56]. Shape parameters corresponding to these measures can be readily obtained by tracking the lip contour and projecting the lip outline onto a recognition basis consisting of the desired lip deformations or attributes. (These steps are referred to as "Track Lips" and "Extract Recognition Features" in figure 5.1.)

In the recognition experiments presented in chapter 6, shape parameters are used as the visual recognition features; however, having tracked the outer and/or inner lip contours, it is straightforward to grab the pixel intensity region bounded by the tracked contours. These pixel values can be used directly as recognition features using principal component encoding [26, 17], or more advanced operators can be applied to make particular judgements about the presence and positioning of the teeth and tongue. Thus, this framework extends naturally to the use of visual features containing both shape and region intensity information.

5.1.4 Audio-Visual Integration

There is some debate within the speechreading community as to the most appropriate time to integrate the audio and visual channels. Early integration, where the audio and visual

feature vectors are concatenated to form one large vector, represents the most natural architecture. Further, since the early integration approach enforces the synchrony of the audio and visual channels, it is well-suited for exploiting bi-modal aspects of the audio-visual speech signal such as voice-onset-time. In addition, the early integration approach is the most general. Indeed, if late integration proved to be the optimal time to integrate, an appropriate learning algorithm should, in theory, learn to treat the channels separately. Accordingly, an early integration approach was adopted here, although during recognition it was possible to vary the weight (importance) of the individual audio and visual components (figure 5.2).

5.1.5   Training

Training for DTW systems reduces to finding which template or templates to use as reference templates for the recognition pattern matching. Several solutions exist for this problem. One approach is to arbitrarily choose one sequence of the concatenated audio and visual features for each word in the vocabulary. Another is to perform clustering analysis on the sequences to generate prototypical reference patterns. The first approach is a little too simplistic and may result in poor reference patterns that do not adequately reflect the data, and the second approach is computationally expensive. A compromise was chosen.

Two exemplar sequences were selected for each word in the database after additional processing. To account for the disparate scales between the audio and visual features and to ensure equal contributions to the Euclidean distance measure used in the recognition stage, the recognition features were normalised to zero mean and unity variance. However, in order to compensate for the fact that the acoustic data contained an unknown and variable amount of noise, the 8 mel-scale acoustic coefficients were normalised over each frame sequence rather than over the training set. This had the undesirable effect of amplifying the noise in clean conditions, although several benefits resulted. Firstly, amplitude normalisation permitted the use of only one set of reference templates for all acoustic noise levels. Secondly, since it is the overall *shape* of the frequency spectrum that identifies the phonemes within the word, not the actual magnitude, normalised data is inherently more robust to changes in volume and noise levels than non-normalised data.

5.1.6   Recognition

Unknown words were recognised by comparing them to the stored reference patterns created in the pre-processing and training phases. (A block diagram of this is shown in figure 5.2 on page 63.) The distance, or dissimilarity, between the unknown utterance and each exemplar

template was calculated using the dynamic time warping algorithm. The template resulting in the minimum cumulative distance was identified as the recognised word.

The DTW algorithm chooses the "best" alignment between two sequences by minimising their cumulative distance. Finding the best alignment between two sequences is functionally equivalent to finding the best path through a grid mapping one sequence to another. Consider two sequences, $X = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_M)$ and $Y = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_M)$, where $\mathbf{x}_t$ and $\mathbf{y}_t$ are L-dimensional feature vectors. Let $d(x_{t_i}, y_{t_j})$ denote the local distance between $\mathbf{x}$ at time t_i and $\mathbf{y}$ and time t_j, and $D(X_{t_i}, Y_{t_j})$ denote the minimum accumulated distance from (X_1, Y_1) to (X_{t_i}, Y_{t_j}). The local distance, or dissimilarity measure, between individual frames of the sequences need not be *distance* in the mathematical sense [111]; it merely needs to possess the property that *similar* sounds result in small distances, while dissimilar sounds produce large distances. In this research a spectral weighted Euclidean distance metric was used,

$$d(x_{t_i}, y_{t_j}) = \sqrt{\sum_{k=1}^{L} w_k (x_k(t_i) - y_k(t_j))^2} \tag{5.2}$$

where w_k is a weighting function used to give greater or lessor importance to the different features.

For recognition using only a single channel w_k was unity for all k resulting in a simple Euclidean distance measure. For recognition using the combined audio-visual data w_k was varied in order to alter the relative weighting between the audio and visual channels.

The total dissimilarity $D(X, Y)$ was computed by finding the minimum cumulative distance over all possible paths from (X_1, Y_1) to (X_M, Y_N), subject to the warping function pair $(\phi_x(k), \phi_y(k))$, and the slope weighting function $m(k)$. Specifically,

$$D(X_M, Y_N) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)) m(k), \tag{5.3}$$

where T is an artificial time reference. The functions $\phi_x(k)$ and $\phi_y(k)$ are monotonically non-decreasing as they must preserve the temporal ordering of the speech signal. In this work, only horizontal, vertical, and diagonal movements were permitted with corresponding slope weights, m_{10}, m_{01}, and m_{11}.

$$D(X_{t_i}, Y_{t_j}) = \min \left\{ \begin{array}{l} D(X_{t_{i-1}}, Y_{t_j}) + m_{10} d(x_{t_i}, y_{t_j}) \\ D(X_{t_i}, Y_{t_{j-1}}) + m_{01} d(x_{t_i}, y_{t_j}) \\ D(X_{t_{i-1}}, Y_{t_{j-1}}) + m_{11} d(x_{t_i}, y_{t_j}) \end{array} \right\} \tag{5.4}$$

Values for m_{10}, m_{01}, and m_{11} were chosen to favour the diagonal path (minimal temporal distortion), while still permitting the horizontal and vertical movements needed to compensate for the simple endpoint detection method employed. The cumulative distance $D(X, Y)$

was normalised by the total path distance $\sqrt{M^2 + N^2}$ to enable comparison of unequal length words.

5.2 Hidden Markov Model Recogniser

The second audio-visual recogniser developed used Hidden Markov Models to represent the words in the vocabulary. Good reviews of basic HMM techniques as applied to speech recognition can be found in [112, 110, 70, 111]. The salient feature of the HMM paradigm is its modelling of the temporal variations in speech signals using statistical methods. As discussed earlier, the speech signal is assumed to be stationary over a short time interval, called the frame interval. The time varying nature of the signal is represented as a concatenation of many short-time stationary segments. The overall speech signal is thus modelled as a synchronous sequence of symbols. These symbols correspond to the audio and visual recognition features (mel-scale filter-bank coefficients and visual shape parameters).

In HMM-based recognition, the symbols are assumed to be generated by a first order Markov process. A first order Markov process is a finite state machine that changes state every time-unit equal to the frame interval, where transition from one state q_t to the next q_{t+1} depends only on the current state. Specifically, transition from state i to state j is probabilistic and governed by the state transition matrix $\mathbf{A} = \{a_{ij}\}$. At each time instance t, statistically independent observation vectors $\mathbf{o}_t$ are generated with probability density $b_j(\mathbf{o}_t)$. (Technically, for speech signals this is an invalid assumption, as speech observation vectors are inherently dependent; however, much success has been achieved using HMMs on speech recognition problems.) The speech signal is considered to be the feature vector sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \ldots, \mathbf{o}_T)$, generated by a given state sequence $\mathbf{q} = (q_1, q_2, q_3, \ldots, q_T)$. In practice, only the observation sequence $\mathbf{O}$ is known, while the underlying state sequence $\mathbf{q}$ remains unknown or *hidden*. Since the stochastic state sequence is only observable through the probabilistic output symbols ($\mathbf{O}$), HMMs are typically referred to as doubly stochastic processes.

Creating a recognition system based on HMMs can be broken into several steps:

- Choose the desired recognition unit (phoneme, tri-phone, syllable, or word). Words were used in the research presented here.

- *Train* an HMM for each recognition unit (word) in the vocabulary using a set of labelled (known) reference utterances.

- Given an unknown observation sequence, calculate the probability that the sequence

was generated by each of the word models.

- Identify the unknown as the model yielding the highest probability.

Not listed above, but implied are the standard tasks of endpoint detection and feature extraction necessary for representing the audio-visual speech signal as sequences of output symbols. This processing was identical to that accomplished in the DTW system and was presented in sections 5.1.1, 5.1.2 and 5.1.3.

A block diagram of the training and recognition sub-systems are shown in figures 5.3 and 5.4, respectively. These two sub-systems are described next.

5.2.1   Training

The most difficult task in HMM-based recognition is training, that is, estimating the parameters for each model given a set of labelled, reference tokens. It is desirable that the trained model be a generalisation of all of the occurrences of the given word. For this research, a commercial software toolkit, The Hidden Markov Model Toolkit (HTK) by Entropic Research Laboratory, Inc., was used to facilitate the building, training, and the manipulation of the continuous density models. An overview of the toolkit can be found in [141].

The HMMs used were characterised by 5 parameters, N, $\mathbf{A}$, B, Π, M:

- N, the number of states in the models

- $\mathbf{A} = \{a_{ij}\}$, the state transition matrix, the probability of moving from state i to state j

- $B = \{b_j(\mathbf{o}_t)\}$, the emission probability vector, the probability of observing $\mathbf{o}$ at time t when in state j

- $\Pi = \{\pi_i\}$, the initial state distribution, the probability of starting in state i

- M, the observation symbols, a continuous quantity represented as a mixture of Gaussians.

In this work, six-state word models were used ($N = 6$). This is consistent with other isolated word recognisers [111] where 5-state models were used for recognition on the digits and [88] where error rates where essentially constant for $N \geq 6$. An example model is shown in figure 5.5. Note that the model actually contains 8 states as opposed to 6. This is a result of the manner in which the HMMs are represented in HTK. All entry and exit states are non-emitting (no observation symbol is generated) with an automatic ($\Pr(a_{ij}) = 1$)
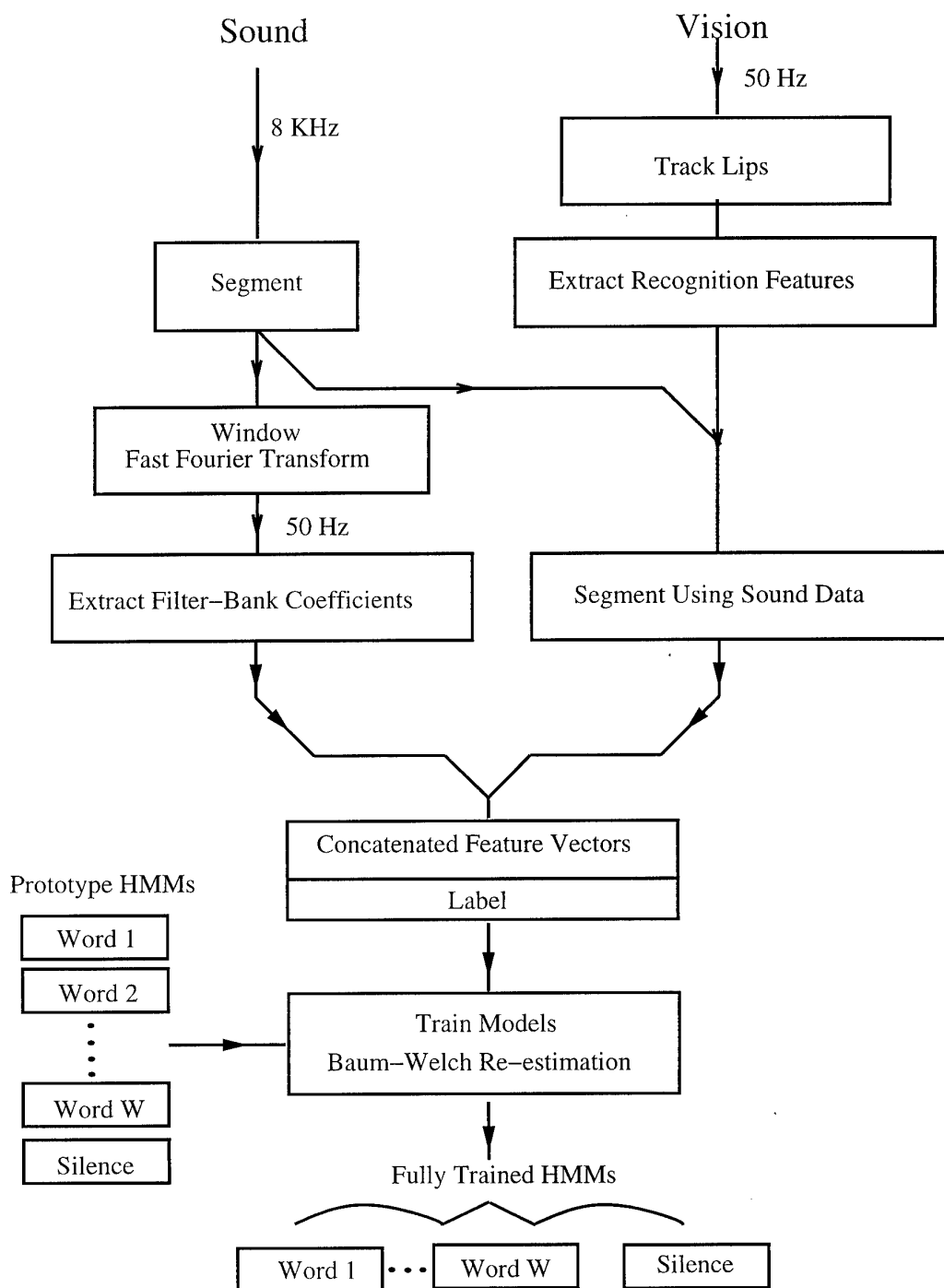
Figure 5.3: *Block diagram of the HMM training sub-system. The raw audio and visual data are processed and relevant recognition features are extracted and used to train continuous density HMMs. A separate model is used for each word (including a "silence" model.)*
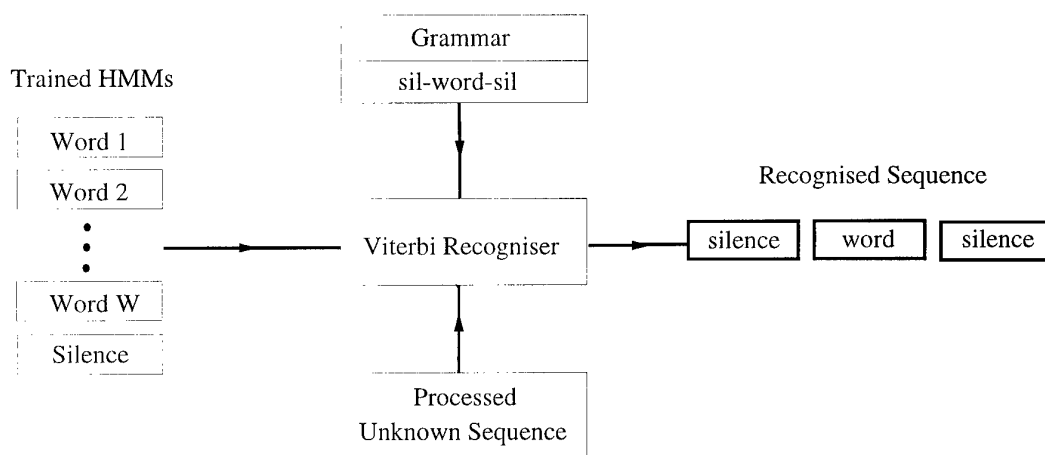
Figure 5.4: *Block diagram of the HMM recognition sub-system. Unknown words are processed and the posterior probability of the unknown sequence, given each of the trained models, is evaluated using the Viterbi algorithm. The "silence-word-silence" grammar forces the recogniser to determine the word endpoints. The word model yielding the highest probability is identified as the "recognised" word.*
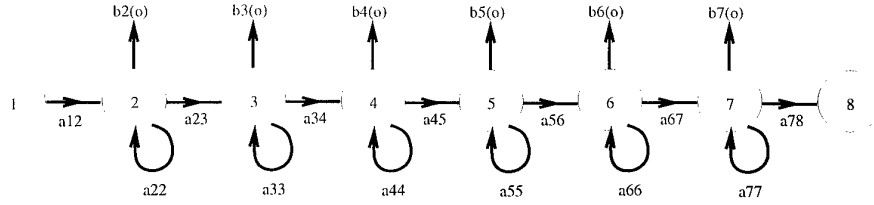
Figure 5.5: *A typical word model used in this research. The transition probability from state i to state j is denoted a_{ij}. The emission probability of observing o in state j is represented as $b_j(\mathbf{o})$. Note that the model contains 8 states, although the first and last are non-emitting resulting in only 6 emitting states. This left-to-right model encapsulates the temporal ordering of the audio-visual data.*

transition from state 1 to state 2 and from state $N - 1$ to state N. Note further, that the model is a left-to-right model. This encapsulates the temporal ordering of the slowly time-varying nature of speech. In theory, each of the states represents a different portion of the speech signal where the vocal tract is essentially static and the transitions from one state to the next represents changes in voicing or repositioning of the articulators.

The model for the "silence" token used $N = 1$. This follows from the belief that during periods of the silence the energy in the signal is a result of background, microphone, or tape noise, which is random in nature and not representative of any words in the vocabulary.

Rather than using codebooks, where M is the number of observation symbols and the emission probabilities, discrete entities, continuous observation densities were used. To accomplish this, the emission probability density function (pdf) was represented as a mixtures of Gaussians. Specifically,

$$b_j(\mathbf{o}) = \sum_{k=1}^{M} c_{jk} \mathcal{N}(\mathbf{o}, \mu_{jk}, \Sigma_{jk}), \quad 1 \leq j \leq N \tag{5.5}$$

where M is the number of mixtures, c_{jk} is the mixture coefficient for the kth mixture in state j, and $\mathcal{N}$ is a standard multi-variate Gaussian with mean vector μ_{jk} and covariance matrix Σ_{jk},

$$\mathcal{N}(\mathbf{o}, \mu_{jk}, \Sigma_{jk}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_{jk}|}} e^{-\frac{1}{2}(\mathbf{o} - \mu_{jk})^T \Sigma_{jk}^{-1} (\mathbf{o} - \mu_{jk})} \tag{5.6}$$

with n the dimensionality of o. The mixture gains c_{jk} satisfy the stochastic constraint

$$\sum_{k=1}^{M} c_{jk} = 1, \quad 0 \leq c_{jk} \leq 1, \quad 1 \leq j \leq N$$

such that the pdf is properly normalised

$$\int_{-\infty}^{\infty} b_j(\mathbf{o}) d\mathbf{o} = 1, \quad 1 \leq j \leq N.$$

Because of the availability of only a limited number of training tokens, the observations in $\mathbf{o}$ were assumed to be uncorrelated resulting in a diagonal Σ_{jk} and substantially reducing the number of free parameters in the model. The audio-only recogniser used a single multi-variate Gaussian to represent the emission densities, while the visual-only and combined audio-visual recognisers used multiple mixture densities.

Having chosen model parameters N and M, the only remaining parameters to estimate were the probability measures $\mathbf{A}$, B, and Π. For ease of notation, let $\lambda=(\mathbf{A},B,\Pi)$ denote the parameter set of the model. Training thus reduces to finding the most likely λ's, in a probabilistic sense, for each of the words in the vocabulary. In essence, for an observation sequence $\mathbf{O} = (\mathbf{o}_1,\mathbf{o}_2,\mathbf{o}_3,\ldots,\mathbf{o}_T)$, one wishes to find values for λ such that the $\Pr(\mathbf{O}|\lambda)$ is maximised. Unfortunately, an analytic solution to this problem does not exist. There does, however, exist an iterative procedure for choosing the maximum likelihood (ML) estimate for the model parameters using an Expectation Maximisation (EM) algorithm. This method is known as the Baum-Welch algorithm [8], and was the method used to estimate the model parameters. A thorough explanation of Baum-Welch estimation using the computationally efficient forward-backward algorithm is given in [70, 111].

5.2.2 Recognition

Recognition of an unknown word represented by a sequence of output symbols $\mathbf{O}$ was accomplished by evaluating $\Pr(\mathbf{O}|\lambda^i)$ (where λ^i represents the model of the ith word in the vocabulary) for each word in the vocabulary and identifying the model yielding the highest probability as the spoken word. These output probabilities can be efficiently calculated using the forward-backward algorithm used in Baum-Welch estimation and is typically the method of choice for isolated-word recognisers. However, despite recognising only single word utterances, a connected-word recogniser was used in this research treating each utterance as a three "word" sequence (silence-word-silence). This was done to compensate for inexact endpoint detection due to the "noisy" speech data and also to serve as a bridge to the more complicated task of continuous-speech recognition. A block diagram of this is shown in figure 5.4.

Thus, instead of computing $\Pr(\mathbf{O}|\lambda)$ by summing $\Pr(\mathbf{O},\mathbf{q}|\lambda)$ over all possible state sequences $\mathbf{q}$, the maximum of $\Pr(\mathbf{O},\mathbf{q}|\lambda)$ (P^*) was computed over the most likely state sequence (q^*) using a dynamic programming technique referred to in the literature as the Viterbi algorithm [138]. The Viterbi algorithm is very similar to dynamic time warping algorithm discussed in section 5.1.6. $P^*(\mathbf{O},\mathbf{q}|\lambda)$ and $q^*(\mathbf{O}|\lambda)$ were computed using the Viterbi algorithm with the model yielding the highest probability being identified as the

recognised word.

5.3   Summary

This chapter has provided an overview of the various sub-systems comprising two audio-visual speech recognisers. These recognisers were developed in order to assess the ability of visual information extracted from the lip contour to provide robust speech recognition. The first recogniser is a dynamic time warping-based isolated word recogniser and the second a connected-word, Hidden Markov Model-based recogniser. Both systems utilise the dynamic contour trackers presented in the previous chapter. Various recognition experiments were conducted using these recognisers and are presented in the next chapter.

6

Audio-Visual Speech Recognition

In order to achieve real-time tracking performance, it is often necessary to reduce the dimensionality of the image data, such as through parameterisation of the lip outline (or outlines) as was done here. Unfortunately, such parameterisation may result in the loss of important recognition information, such as the position of tongue and teeth. This loss of information is of special concern as some researchers [17, 19] have concluded that the outline of the lip is not sufficiently distinctive to give reliable recognition performance. One aim of this thesis is to demonstrate that even partial visual information, in the form of *shape* parameters describing the lip outline, can enhance speech recognition. This chapter presents experiments to test if this is indeed the case. Particular emphasis is placed on the *incremental vision rate*, which is the increase in recognition performance due to the incorporation of visual information into the acoustic speech recognisers. Essentially, it is a measure of the additional recognition information provided by the visual data.

The DTW and HMM recognisers described earlier (sections 5.1 and 5.2) are used for these experiments. Recognition performance is assessed using isolated-word vocabularies with and without added Gaussian noise. Tests are accomplished using audio-only, visual-only, and combined audio-visual data. The visual information is represented as projections of the lip outline onto three different recognition bases. The bases examined are the affine basis and two bases learnt from principal components analysis. First, the method used to determine these recognition bases is presented, followed by the results of the recognition

experiments.

6.1 Lip Motion and Visual Feature Extraction

Prior to presenting the results of the recognition experiments, a detailed analysis of the characteristic lip motions found in natural speech is given. This analysis serves two purposes. First, it provides insight into the dominant lip movements present in visual speech. Second, it permits compact representation of lip shape in terms of basis vectors characteristic of natural lip movements, which were then used in the recognition experiments.

6.1.1 Shape Models for Lip Deformations

In order to do justice to the complex articulatory movements of the lips in natural speech, it is necessary to parameterise their outline with many control points (eg. 11 (x, y) coordinate pairs or $N_X = 22$ degrees of freedom). However, as discussed in section 3.3, it is desirable to limit the number of degrees of freedom of deforming lips during tracking, both for stability and for computational reasons. This can be accomplished by imposing shape constraints on the allowable lip deformations. Such restrictions on the possible space of lip shapes seem natural as lip movement can be largely accounted for by just a few independent modes of motion. The question becomes how to best choose this shape space.

One possible method is to hand-fit contours to a representative set of lip deformations, the so-called "key-frame basis". However, such construction is largely an art and there is no guarantee that the basis vectors chosen will be optimal or even that they will span a majority of the lip deformations found in natural speech. A second alternative is to hand-fit splines to a large number of lip configurations and then perform a data reduction technique such as principal components analysis [31] to the hand-fit sequence. Such an approach can be quite effective and is precisely the method employed by researchers using Active Shape Models [83, 86, 87]. A third, somewhat preferable, method is to *learn* the space of lip deformations automatically from tracked sequences of articulating lips. Naturally, it is first necessary to be able to track the lips.

The cosmetically-assisted lip tracker, discussed in section 4.3, permitted tracking of the lips in high-dimensional spaces with a high degree of accuracy. In order to capture the full range of mouth shapes present in natural speech, a single speaker uttering a continuous sequence of words containing the 40 American English phonemes was recorded onto video tape (60 seconds = 3000 fields). The speaker's lips were then tracked throughout this sequence and the lip deformations corresponding to this speech were stored in control point

form. Principal components analysis was then performed on the extracted data.

6.1.2 Principal Components Analysis using the L_2-norm

Principal components analysis (PCA) is a proven method for determining the principal axes of variation in a data set. Thus, it can be used to determine the main modes of lip deformation in natural speech, as well as the degree to which each component contributes to the variation in the entire space of lip deformations. The results of the principal components analysis can also be used to provide possible bases for recognition.

The training data were recorded in control point space, when actually it is the lip deformations in *spline* space that are of interest. Therefore, PCA was done using the L_2-norm and inner product as defined in equations 3.2 and 3.4. Using this norm, it can be shown [31] that the principal components in spline space are given by the eigenvectors $\mathbf{v}_i$ of

$$\sum_k (\mathbf{X}_k - \overline{\mathbf{X}}')(\mathbf{X}_k - \overline{\mathbf{X}}')^T \mathcal{H} \tag{6.1}$$

with corresponding eigenvalues λ_i, where the $\mathbf{X}_k$ are the sequences of control point vectors and $\overline{\mathbf{X}}'$ is the mean of the sequences. The proportion of lip motion variance accounted for by a particular eigenvector $\mathbf{v}_i$ is given by

$$\frac{\lambda_i}{\text{Trace } \Lambda}, \tag{6.2}$$

where Λ is a diagonal matrix of the eigenvalues.

Essentially the PCA problem is, given a training sequence $\mathbf{X}_1, \ldots, \mathbf{X}_M$, find a template vector $\mathbf{Q}'_0$ and shape matrix $W' = (\mathbf{v}_1, \ldots, \mathbf{v}_{N'_Q})$ such that the reconstructed sequence $\mathbf{X}'_1, \ldots, \mathbf{X}'_M$ most closely approximates the training sequence in a least-squares sense:

$$\min_{\mathbf{Q}'_0, W'} \left(\sum_{k=1}^M \|\mathbf{X}_k - \mathbf{X}'_k\|^2 \right), \tag{6.3}$$

where

$$\mathbf{X}'_k = W'\mathbf{Q}'_k + \mathbf{Q}'_0.$$

The resultant shape space $\mathcal{S}_{Q}' = \mathcal{L}(W', \mathbf{Q}'_0)$ is then spanned by the basis vectors $\mathbf{v}_i$ which are the columns of the shape matrix W plus an offset template $\mathbf{Q}'_0$. The complete algorithm is described in figure 6.1.

When the training sequence is obtained by hand-fitting splines to the lip contours or running a tracker which utilises the maximum number of degrees of freedom available (N_X = twice the number of control points), Σ has rank N_X (provided $M \geq N_X$). However,

---

Principal Components Analysis

Given: Training control point data $\mathbf{X}_1, \ldots, \mathbf{X}_M$ in spline space $\mathcal{S}$

Find: A sub-space $\mathcal{S}_Q' = \mathcal{L}(W', \mathbf{Q}_0')$ of dimension N_Q' to minimise

$$\sum_{k=1}^{M} \|\mathbf{X}_k - \mathbf{X}_k'\|^2$$

where

$$\mathbf{X}_k' = W'\mathbf{Q}_k' + \mathbf{Q}_0'.$$

Algorithm

1. Construct the training-set mean

$$\overline{\mathbf{X}}' = \frac{1}{M} \sum_{k=1}^{M} \mathbf{X}_k.$$

2. Construct the training-set covariance

$$\Sigma = \frac{1}{M} \sum_{k=1}^{M} (\mathbf{X}_k - \overline{\mathbf{X}}')(\mathbf{X}_k - \overline{\mathbf{X}}')^T \mathcal{H}.$$

3. Find eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_{N_Q'}$ of Σ, in descending order of eigenvalue.

4. The principal components parameters $\mathbf{Q}_0'$, W' of the shape sub-space are then

$$\begin{aligned} \mathbf{Q}_0' &= \overline{\mathbf{X}}' \\ W' &= \left(\mathbf{v}_1, \ldots, \mathbf{v}_{N_Q'}\right). \end{aligned}$$

Figure 6.1: *Algorithm for L_2 PCA in spline space.*

when the training data are obtained automatically by tracking real lip sequences, it is often necessary to track in a reduced shape space, $\mathcal{S}_Q = \mathcal{L}(W, \mathbf{Q}_0)$, with $N_Q < N_X$ degrees of freedom, to prevent tracker instability. In such cases, the tracking shape matrix W can be built using key-frames as discussed in section 3.3 with the only stipulation being that the resultant tracking space $\mathcal{S}_Q$ adequately span the lip deformations found in natural speech. The principal components analysis algorithm (figure 6.1) is equally suitable for data gathered in this manner, although the control point sequence $\mathbf{X}_1, \ldots, \mathbf{X}_M$ is replaced by $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_M$ where

$$\tilde{\mathbf{X}}_k = W\mathbf{Q}_k + \mathbf{Q}_0$$

to indicate the reduced degrees of freedom of the training data. For the analysis presented here, a 10 degree of freedom shape matrix was used to track the 3000 field sequence of continuous speech.

The percentage of lip deformations explained by the most significant principal component is shown in table 6.1, where it is seen that the first principal component accounts for over half of all lip movements, the first three 94%, and the first six 99%.

Component	1	2	3	4	5	6	7	8	9
Mean Square Lip Movement	52.3%	23.4%	18.4%	3.1%	1.6%	0.4%	0.3%	0.2%	0.2%
Cumulative Sum	52.3%	75.7%	94.1%	97.2%	98.8%	99.2%	99.5%	99.7%	99.9%

Table 6.1: *The percentage of lip deformation variance explained by each of the first 9 principal components. Note that the first component accounts for over half of all lip movement variance and that 94% of all lip motion can be expressed with only the first 3 principal components, and 99% with the first 6.*

An attractive feature of the dynamic contour tracking framework is that the actual lip motion represented by a given basis vector can be observed by converting the dominant eigenvector into vibration modes of the control points of the lip template. Table 6.2 and figure 6.2 show the direction and relative magnitude of the first principal component and its motion superimposed on the lip template.

	Control Point Number										
	1	2	3	4	5	6	7	8	9	10	11
Angle	−84°	−120°	−136°	−138°	−94°	−64°	−108°	−79°	−90°	−106°	−109°
Size	13.7	5.3	4.4	0.7	2.6	2.6	11.0	18.3	23.5	24.9	15.8

Table 6.2: *The dominant mode of vibration generated by the dominant eigenvector for each control point in terms of magnitude and direction. Figure 6.2 shows these motions superimposed on the lip template.*
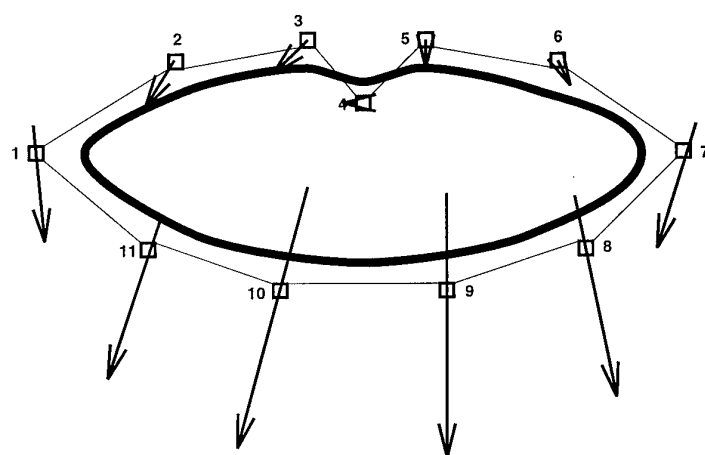
Directions of Main Eigenvector
Size Indicates Relative Strength

Figure 6.2: *The dominant lip motion present in normal speech. This is clearly vertical displacement of the lower lip caused by opening of the mouth.*
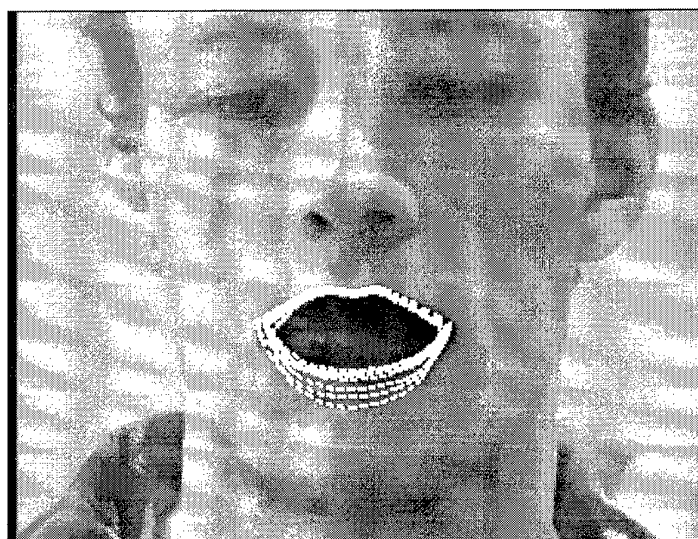
Figure 6.3: *Multiple traces of the lip tracker superimposed on a static image of a speaker demonstrating that most lip deformation can be explained by vertical displacement of the lower lip. (Figure courtesy of Barney Dalton.)*

By observing the dominant mode of vibration, it is clear that vertical displacement of the lower lip (caused by opening of the mouth) is the prominent lip movement in spoken English. This observation is also evident in figure 6.3 where multiple traces of the lip tracker are superimposed on the speaker. This finding should not be surprising as the primary motion of speech is the lowering of the jaw, which moves concurrently with the lower lip.

It is also instructive to look at the remaining principal components expressed as deformations of the basic lip template — that is, the lip movements they represent. Figure 6.4 shows the deformations along each of the first six principal components axes in terms of their shape and significance. As seen earlier, the first component represents the degree of "lip opening" due to vertical displacement of the lower lip and accounts for the majority of the variance in lip movements. The second component appears to represent a combination of global movement and a slight bit of vertical scale (ie. lip opening). This highlights one of the classic problems faced by all automatic speech readers, that is, the coupling of the "local" lip/mouth movements and the "global" head or body movements. This is especially important given that the vertical displacement of the lower lip could be the result of a speaker opening his/her mouth, nodding his/her head, or standing up/squatting down. It is true that each of these movements represents a slightly different deformation (ie. opening of the mouth is vertical displacement of the lower lip only, nodding of the head is vertical displacement of the lower lip and vertical shrinkage of the template, and squatting down
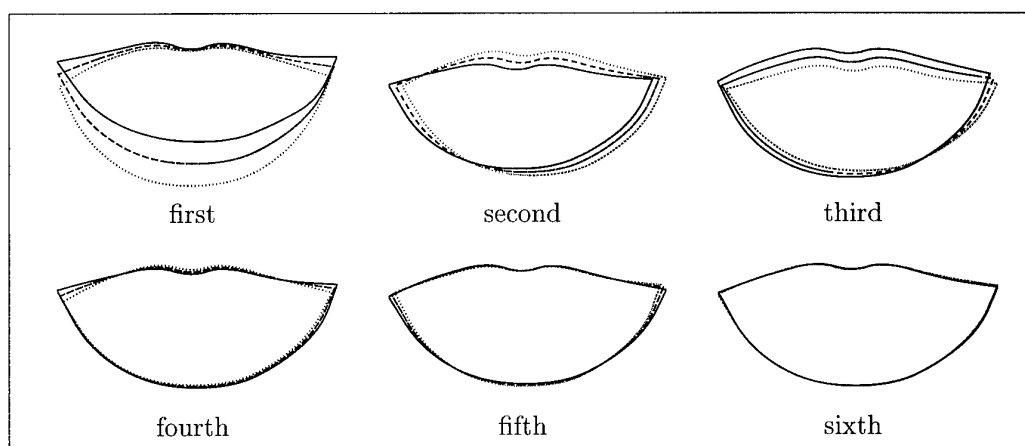
Figure 6.4: *The lip deformations corresponding to the* **principal components** *plotted two standard deviations either side of their mean. The first component represents the degree of lip opening resulting from movement of the lower jaw. The second component is a combination of horizontal and vertical movement of the template and a small amount of vertical scaling. The third component is curling of the upper lip with some global displacement. The fourth, fifth, and sixth components account for such as a small percentage of lip movement that it is difficult to see what motion they represent.*

is vertical displacement of both the upper and lower lip) but, accomplished in concert, it would be nearly impossible to separate the individual movements using only a single tracker. However, simple global head motion can be decoupled from the articulatory movements of the lip by simultaneously tracking a fixed position on the head such as the nostrils [106], or tracking the entire head [78], and then subtracting these global displacements from the lip control point positions. The decoupling of head pose from lip movement/facial expression for more general head orientations remains an open research area, although results from recent work [10, 6] are promising.

The third component also appears to be a composite of global head movement and local lip deformation with the lip movement being the rounding of the upper lip similar to that seen in a 'pr' sound. The fourth, fifth, and sixth components account for such a small percentage of overall mean square lip movement (less than 6% in total) that it is difficult to see what motion they represent when plotted commensurate with their significance. It is easier to visualise their deformations in figure 6.5 where they are plotted a normalised distance either side of their mean.

The fourth component represents a curling of the lip corners similar to that required to produce the 'ee' sound. The fifth and sixth are harder to classify as they account for only two percent of the lip movements and may not be representative of any particular lip
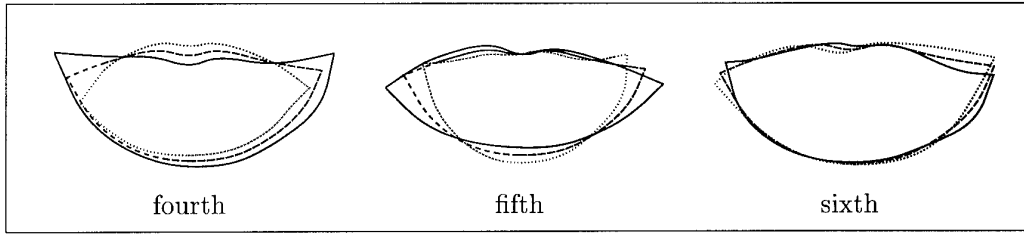
| fourth | fifth | sixth |

Figure 6.5: *The lip deformations corresponding to* **principal components** *four, five, and six plotted a normalised displacement either side of their mean. The fourth component represents curling of the lip corners; the fifth, horizontal scaling and curling of the lip corners, and the sixth, an asymmetrical deformation.*

movement involved in speech production.

The speaker used in this research spoke naturally, limiting unnecessary head movement. However, the above analysis shows that incidental movements were still present — an unavoidable characteristic of human communication patterns. Since it is believed that global horizontal motion is not necessary for speech production and only a by-product of spurious head movements (global vertical displacement is present as a result of the asymmetrical movement of the upper and lower lips), additional analysis was performed on the control points with horizontal displacement (X translation) removed. For each set of control points, the mean horizontal translation component was removed using

$$\tilde{\mathbf{X}}_k = \mathbf{X}_k - \frac{\langle \mathbf{X}_k, \mathbf{X}_T \rangle}{\langle \mathbf{X}_T, \mathbf{X}_T \rangle} \mathbf{X}_T \qquad (6.4)$$

where $\mathbf{X}_T$ is the horizontal displacement (the first vector in the affine basis) and the norm and inner product are defined in (3.2) and (3.4).

Principal component analysis was then performed on the resultant data ($\tilde{\mathbf{X}}_k$) yielding the deformations of figure 6.6. These deformations are subsequently referred to as the "PCA no X" basis. The first four deformations can be broadly classified as

- lower lip movement – 'ah' sound

- rounding of upper lip – 'pr' sound

- curling of lip corners – 'ee' sound

- scaling of lip corners – 'wa' sound

Components five and six are again difficult to classify. Looking at their contribution to the total space of lip deformations (table 6.3) we see that they account for only 1.6% and
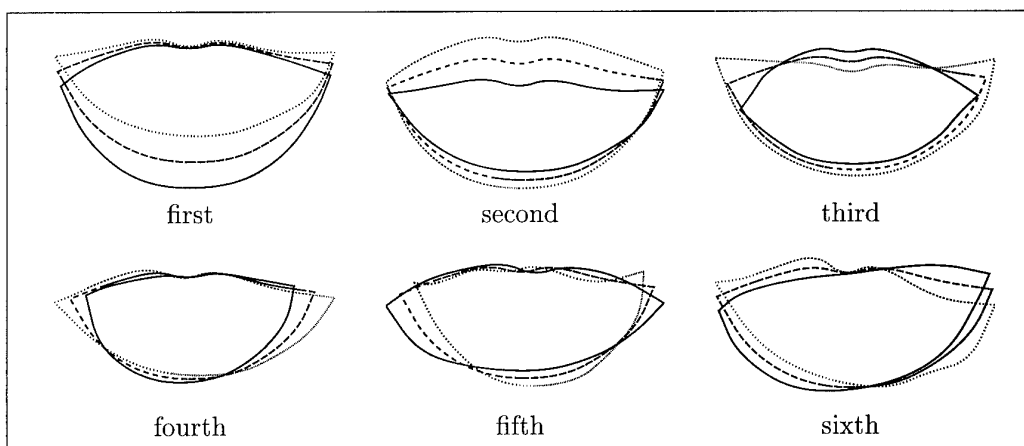
Figure 6.6: *The lip deformations corresponding to the first 6* **principal components** *produced after* **subtracting X translation** *plotted a normalised displacement either side of their mean. The first four deformations can be broadly classified as lower lip movement, rounding of the upper lip, curling of the lip corners, and horizontal scaling.*

Component	1	2	3	4	5	6	7	8	9
Mean Square Lip Movement	65.8%	24.9%	4.2%	2.4%	1.6%	0.4%	0.3%	0.2%	0.1%
Cumulative Sum	65.8%	90.7%	94.9%	97.3%	98.9%	99.3%	99.6%	99.8%	99.9%

Table 6.3: *The percentage of lip deformations explained by each of the first 9 principal components computed after subtracting horizontal displacement from the control points. The first component accounts for over 65% of all lip movements; the first three, 95%, and the first six, 99%.*

0.4% percent of all lip movements and hence may not be representative of any particular lip movement involved in speech production.

Principal components analysis has provided a means for analysing the space of lip deformations present in normal speech and suggests that 99% of these deformations can be represented with as few as 6 free parameters. In addition to providing insight into the primary deformations of articulating lips, the two sets of PCA basis vectors (PCA and PCA no X) also provide a natural means for compactly representing lip shape information for subsequent recognition experiments.

6.1.3 Affine Basis

Although the space of lip motions can be expressed in descending order of power content using the PCA bases, such visual feature representation may not yield the optimum recognition results. For good recognition it is desired to find feature representations which are the most *discriminating*, as opposed to those that account for the largest percentage of

variance. Thus, while representation of the lip deformations using the PCA bases may be optimal for reconstruction, they may not be optimal for discrimination. Ideally, one would like to identify those lip movements most beneficial for speech recognition, possibly using some form of discriminant analysis. However, the non-linear temporal variations inherent in speech recognition differ from the classic static classification problem [46, 111]. In addition, such deformations represent movements of the control points within the lip template and hence are specific to the speaker and the template parameterisation. In order to develop a visual recognition system capable of adapting to various speakers with different mouth shapes and corresponding lip templates, it is necessary to describe lip movements in more universal terms. Thus, in addition to using visual features obtained by projecting the tracked lip outline onto the two PCA bases, a third basis, the affine basis, was also used.

The affine basis, which was discussed earlier as a means of applying shape constraints to the lip deformations, represents a potential speaker-independent basis. Lip deformations are expressed in the affine basis in terms of translation, scaling, rotation, and shearing of the lip template. These deformations, which were initially described in figure 3.2, are reproduced in figure 6.7 for reference.
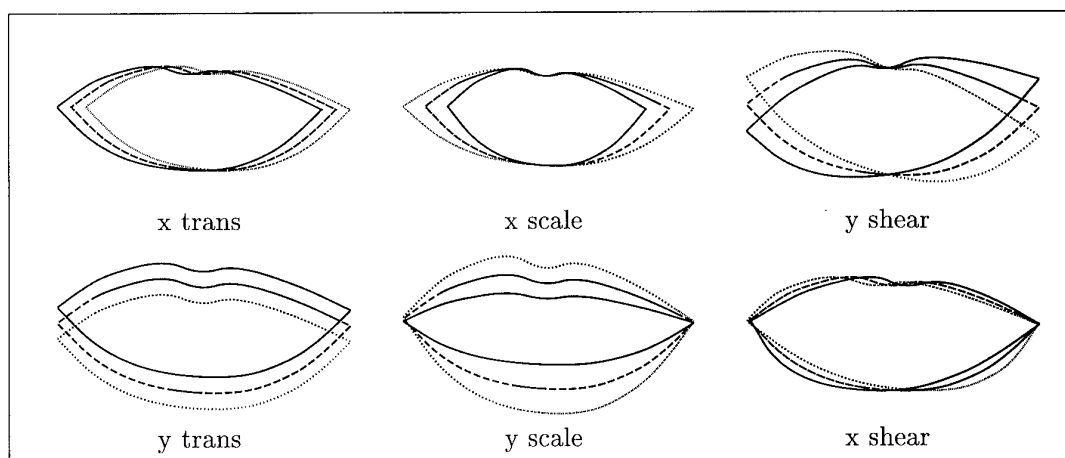
Figure 6.7: *Lip movements corresponding to* **affine** *deformations of the mouth template. The first two components represent horizontal and vertical displacement/translation. The third and fourth, horizontal and vertical scaling, and the fifth and sixth, vertical and horizontal shearing.*

On a sample data set, it was found that the affine basis could account for only 91% of the lip motions as opposed to the 99% accounted for by the PCA bases. Thus, during the recognition experiments it will be important to compare the error rates achieved using the affine basis with those obtained using the more specialised PCA bases and determine to

what extent its generality impairs recognition performance.

6.1.4 Visual Feature Extraction

In order for the visual recognition features to be useful for discrimination, they should (i) be similar across multiple repetitions of the same word, (ii) be sufficiently different between repetitions of different words to provide for linguistic discrimination, and (iii) be indicative of motions characteristic of natural speech. Having decided to represent the lip outline as linear combinations of recognition bases, it is instructive to examine the resultant visual features. It is particularly important to see whether or not the features chosen are repeatable across multiple repetitions of the same word, yet sufficiently different between repetitions of different words, in order to complement the acoustic features.

Traces of the six affine features for multiple repetitions of the words "previous" and "more" are shown in figures 6.8 and 6.9. In both figures we see that components 2 (vertical translation), 3 (horizontal scale), 4 (vertical scale) and 5 (vertical shear) are consistent across all four repetitions. It is not surprising that the X scale and Y scale components are repeatable across repetitions as they represent the crudest indication of overall mouth shape. The repeatability evident in the Y translation component is encouraging. This suggests that the vertical translational component due to incidental head movements was small in magnitude when compared to the translational component resulting from the opening of the mouth. While this may not be true for the entire data set, or for all data sets in general, the preliminary indication is that for this data set the Y translational component was not significantly corrupted by incidental head movements. The consistency in the Y shear component was unexpected (although the recognition experiments confirm that it does in fact contain useful recognition information). The lack of consistency in the X translation and X shear components was expected as neither appears to play a role in the production of speech. In particular, as discussed earlier, it is believed that the X translation component merely reflects spurious head movements of the speaker.

An additional point, not immediately obvious from the graphs but worthy of mention, concerns the lip positioning at the beginning and ending of the words. The speaker started from the rest position (closed mouth) for each utterance. The vision signal begins 100 ms prior to the onset of the audio signal to account for the pre-positioning of the lips, hopefully resulting in an identifiable start for each word. However, it can be seen from the graphs that the starting lip position varied from one repetition to the next, implying that the anticipatory effects of the articulators are not as predictable as was hoped. Furthermore, post articulatory movements (at the end of a word/sentence) are even more variable, as
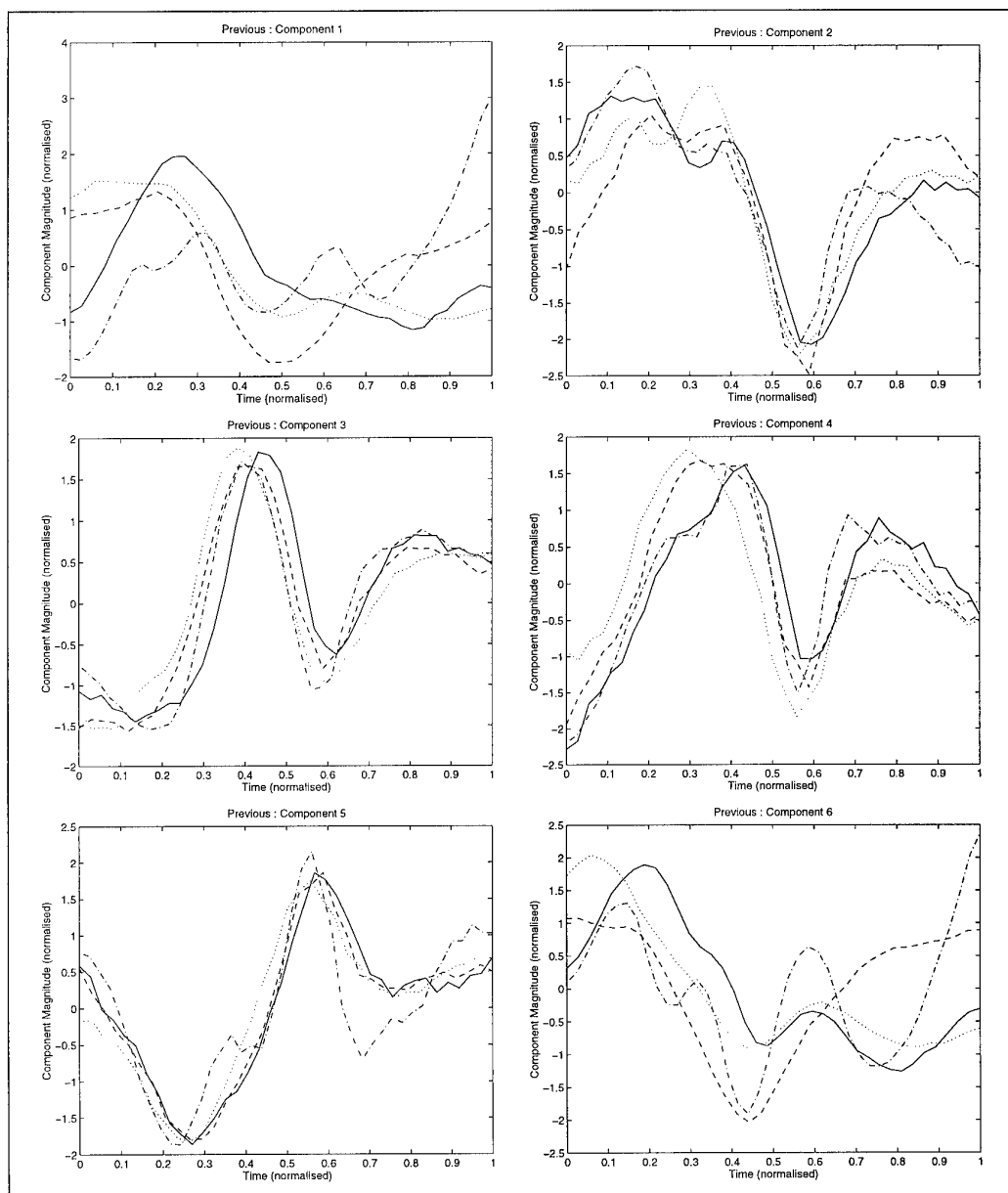
Figure 6.8: *Affine components 1 through 6 for four repetitions of the word "previous". The visual signals for components 2 (Y translation), 3 (X scale), 4 (Y scale) and 5 (Y shear) are similar across all four repetitions which suggests that they may be useful for recognition purposes.*
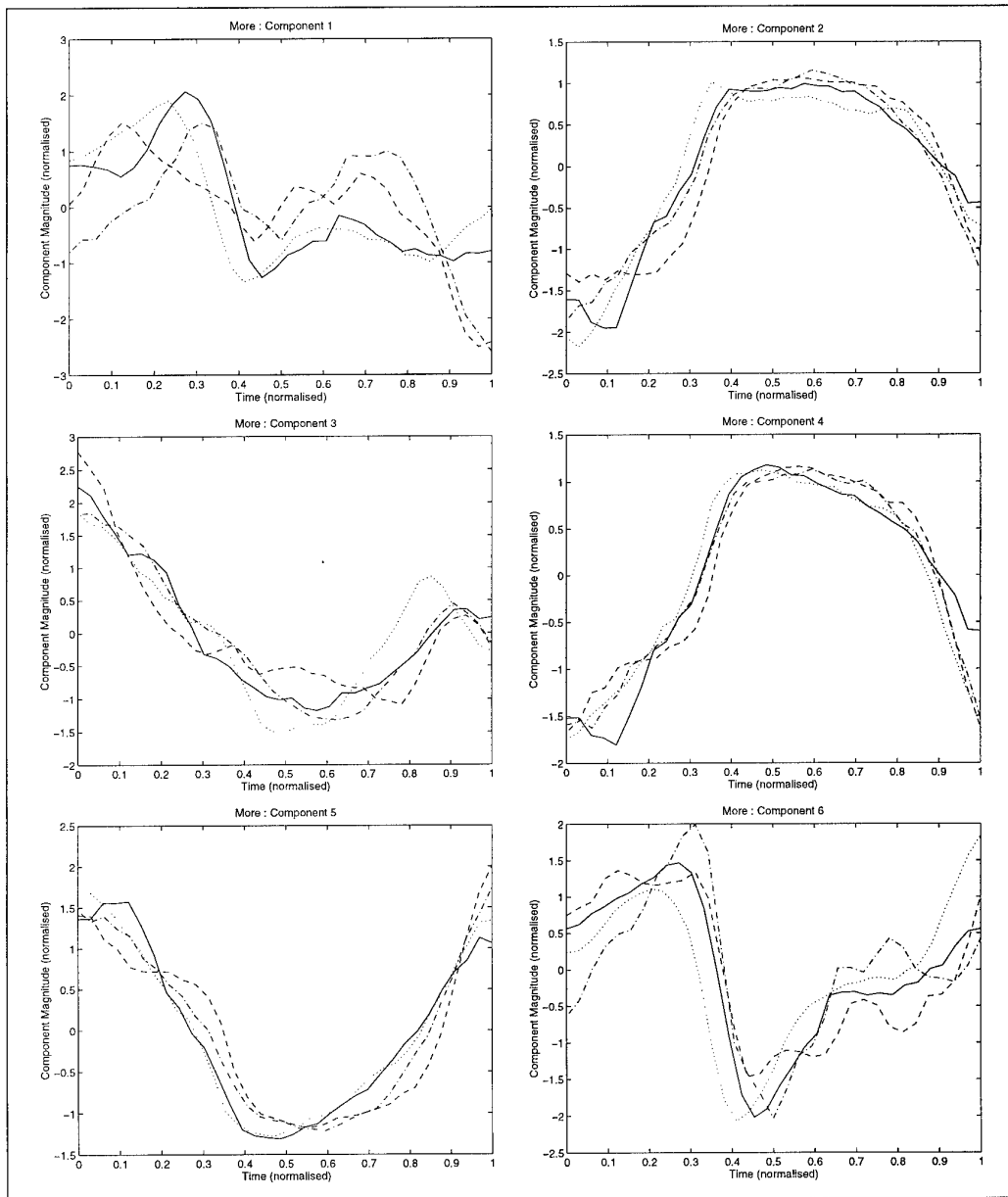
Figure 6.9: *Affine components 1 through 6 for four repetitions of the word "more". Once again, the visual signals for components 2 (Y translation), 3 (X scale), 4 (Y scale) and 5 (Y shear) are consistent across all four repetitions which suggests that they may contain useful recognition information. Furthermore, since they are also significantly different from like features in "previous", they may contain discriminatory information as well.*

with no requirement to pre-position the articulators for the next utterance, the speaker is free to move his lips to any comfortable position. Thus, these graphs serve to illustrate just some of the difficulties in accurately segmenting speech using the visual signal alone.
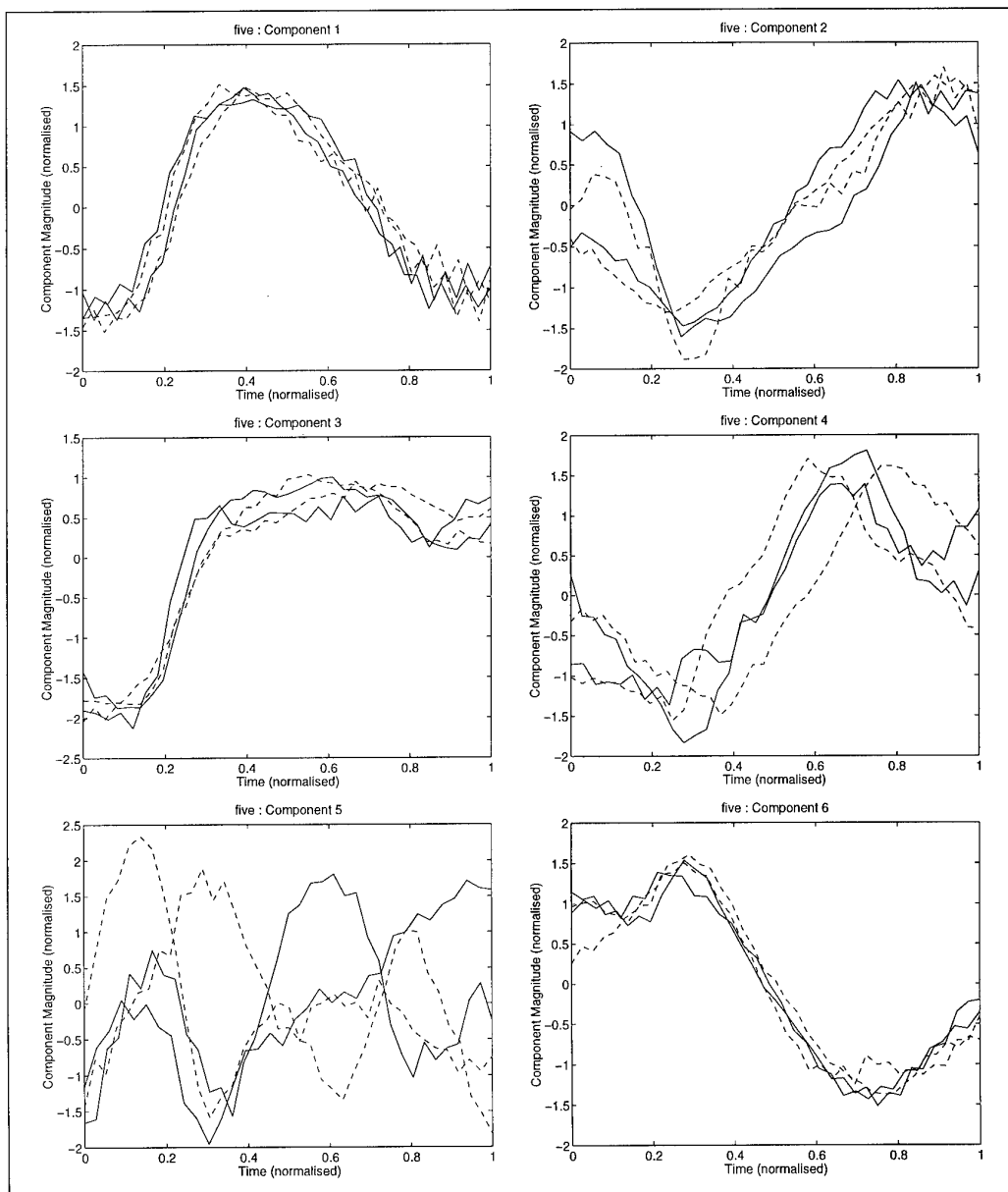
Figure 6.10: *PCA no X components 1 through 6 for four repetitions of the word "five". The visual signals for components 1,2,3,4,6 are similar across all four repetitions. The 1st component, which indicates the degree of mouth opening, reveals "five" as a simple open mouth then close mouth movement.*

Traces of the visual signals for the PCA no X basis, similar to those shown for the affine
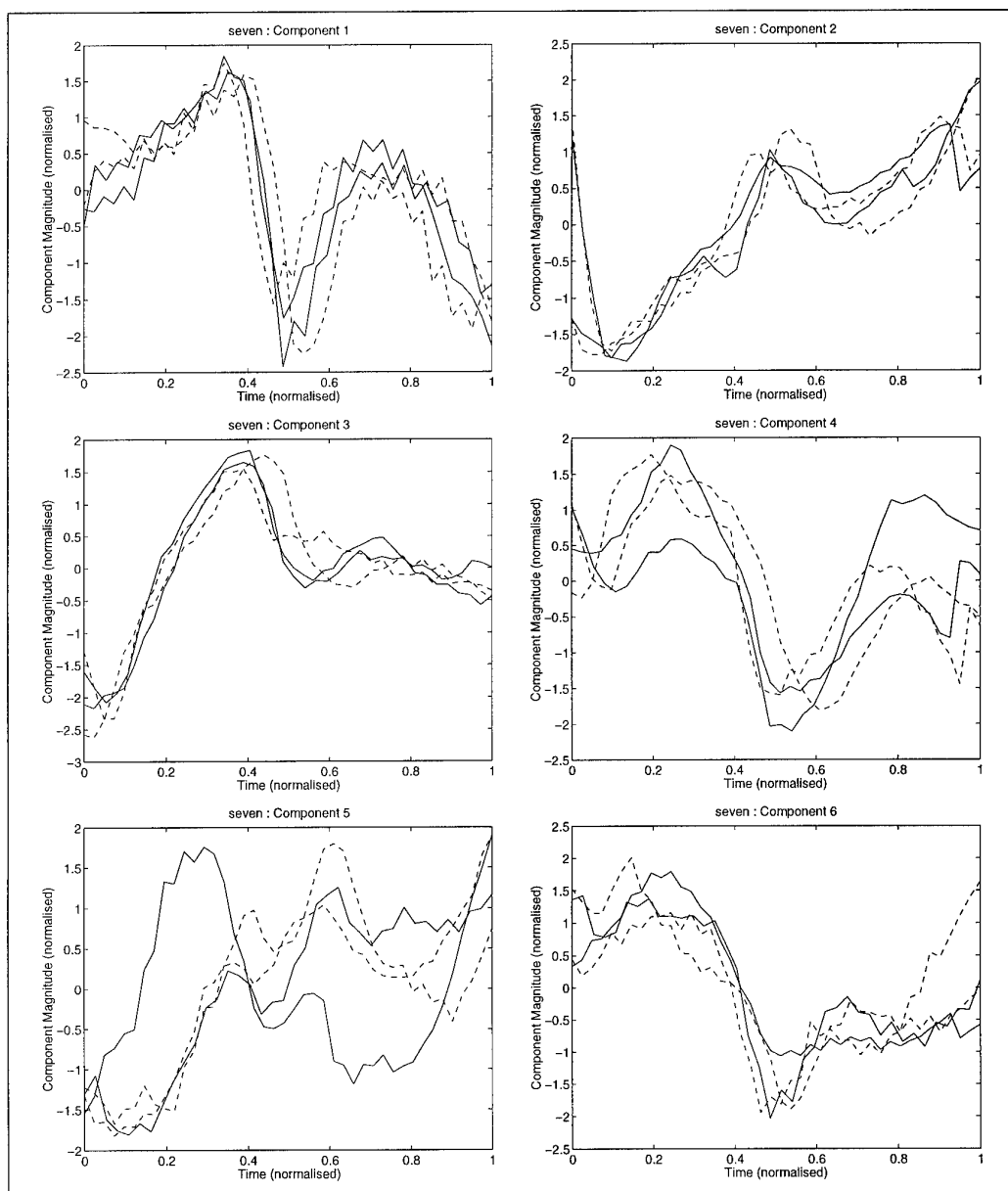
Figure 6.11: *PCA no X components 1 through 6 for four repetitions of the word "seven". The visual signals for components 1,2,3,4,6 are similar across all four repetitions. The rapid motion of the fricative 'v' in the middle of "seven" is clearly seen in the first principal component (degree of mouth opening). This articulatory movement should represent a distinctive characteristic of "seven".*

basis (figures 6.8 and 6.9) are shown in figures 6.10 and 6.11. Once again, several of the visual components are similar across the four repetitions. Furthermore, when comparing the feature traces between different words (eg. "previous" and "more", "seven" and "five"), one sees that the visual signals are substantially different, which is not surprising as their respective articulatory lip movements are markedly different. However, this shows that the shape parameters chosen may encapsulate enough of the linguistically discriminant information to be able to differentiate among words. This would allow features extracted in real time to be used in lipreading applications, paving the way for useable automatic lipreaders.

6.1.5 Summary

Analysis of natural, continuous speech for a single speaker has shown that over 50% of the lip movement present in normal speech is due to the vertical displacement of the lower lip caused by opening of the mouth. The analysis also suggests that it may be possible to express the information content of the lip outline with only a few (≤ 6) free parameters. Visual recognition features can be obtained by representing the lip contour as a linear combination of basis vectors. Three different bases — PCA basis, PCA no X basis, and the affine basis — have been presented and will be used in subsequent recognition experiments.

6.2 Dynamic Time Warping Recognition

6.2.1 Recognition using the Profile View

Although the main experiments were conducted using a frontal view of the face, it seemed important to run at least a pilot experiment using the side-view, given that profile tracking is robust and markedly easier than tracking from frontal views. This experiment was performed to demonstrate that real-time (50 Hz), unaided visual tracking could be useful in audio-visual speech recognition applications, even if only on a modest scale. In place of the 40-word database used for the main experiments, a 10-word digit database was used, containing 20 repetitions of each digit. Rather than exploring alternative representations of the visual information, four simple visual features were used — two affine components and two non-rigid displacements derived from key frames which were believed to convey linguistically pertinent information (lip protrusion and lower jaw movement).

In order to provide an opportunity for the visual signal to contribute to the audio-visual recognition performance, the audio signal was degraded in two ways. Firstly, the speaker varied his distance from the microphone — behaving as if he were speaking into an

automated teller machine, where it is unlikely that all speakers would address the machine from the same distance. Secondly, varying levels of artificial Gaussian noise were added to the acoustic signal corresponding to different signal-to-noise ratios (SNRs).

Audio-only, visual-only, and combined audio-visual experiments were conducted at 4 separate SNRs (clean, 0 dB, -3 dB, and -6 dB). The audio and visual features were combined using an *early* integration strategy. For audio-visual recognition, the relative importance of the two channels was altered by varying the weight of the audio and visual components in the distance metric of the dynamic time warping algorithm (5.2). The resultant error rates for the recognition experiments conducted at the various SNRs with different sound-to-vision weightings are shown in figure 6.12.
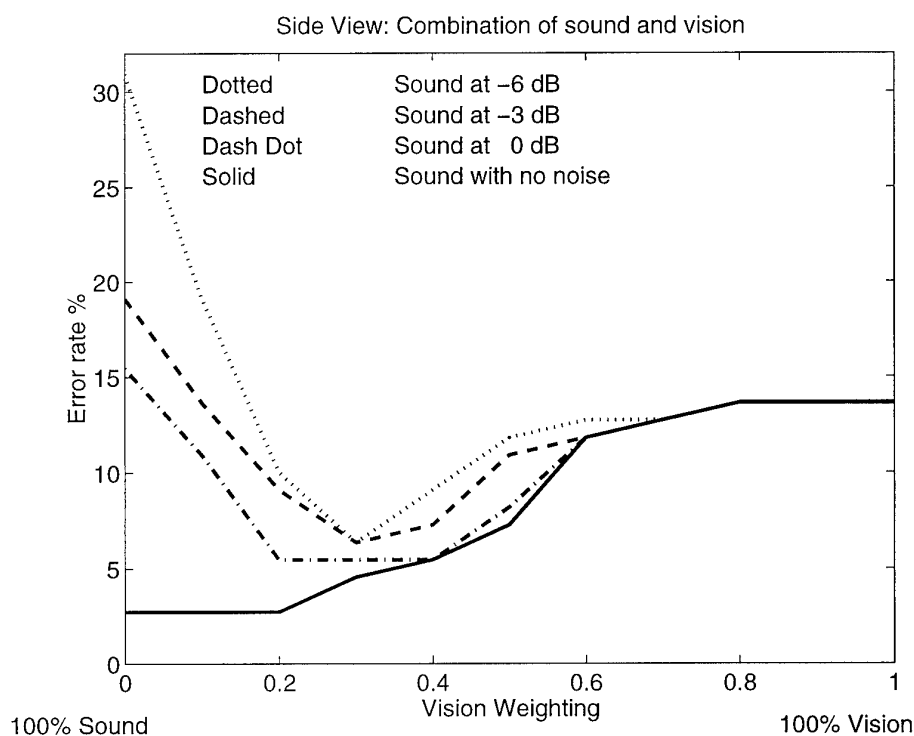
Figure 6.12: *Side View: Error rate variation on the test set as sound-to-vision weighting is varied. Addition of the visual information in the noise-free environment (solid line) does not provide any improvement in recognition performance. However, the incorporation of visual features does improve recognition performance in all three acoustically-degraded scenarios. Further, the combined audio-visual performance exceeds both the audio-only and visual-only performance in all cases. For example, at 0 dB (dashed-dotted line), the error rate using only the acoustic data is 15.5% (far left). The error rate using only the visual data is 13.6% (far right), while combined audio-visual recognition results in an error rate of 5.5%.*

Addition of the visual information does not provide any improvement in recognition

performance in the noise-free environment. However, the incorporation of visual features does improve recognition performance in all three acoustically-degraded scenarios. Not surprisingly, the contribution of the visual information increases with the level of acoustic degradation.

The results of this pilot experiment using the profile lip tracker are encouraging, as they have demonstrated that real-time lip tracking can be used to enhance speech recognition in adverse acoustic environments. However, from a speech recognition standpoint, frontal (or at least partially frontal) viewing is preferable to profile viewing [9], because in profile viewing, the tongue and teeth are often not visible. Furthermore, there may also be a loss of shape information in the lip contour itself, since its width is not directly observable in profile. For these reasons, additional experiments were conducted on a larger 40-word database using the frontal lip tracker (with lipstick).

6.2.2 Recognition using the Frontal View

Isolated-word recognition experiments using the DTW recognisers were conducted on a 40-word database consisting of numbers and commands that might be found in an interactive voice system controlling a car-phone, fax machine, computer, or other office equipment. The words, listed in table 6.4, were carefully chosen to ensure at least one example of each phoneme. Further, numerous similar-sounding words, such as "yes" and "less", "more" and "four", "no" and "show", "two" and "goto", were included to provide a challenging recognition environment. Twenty repetitions of each word were recorded onto video tape and

Forty Word Database									
Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Faster	Slower	Up	Down	Right	Left	Little	Big	More	Less
On	Off	Start	Stop	Clear	Reset	Yes	No	Dial	Hang-Up
Ring	Cancel	Next	Previous	Jump	Goto	Change	Switch	Show	Void

Table 6.4: *A 40 word command-oriented database containing at least one example of each phoneme. Twenty repetitions of each word were recorded.*

partitioned into three sets. Two repetitions were used as exemplar patterns for matching, seven were used as a training set, and eleven as a test set. This resulted in a training set of 280 words and a test set of 440 words.

Raw visual and audio data were gathered simultaneously and in real-time (50 Hz) on a Sun IPX workstation with a Datacell S2200 framestore. The visual data consisted of the mouth outline represented as 13 (x, y) control points and the audio data 8-bit μ-law sampled

at 8 KHz. Acoustic recognition features consisted of 8 mel-scale filter-bank coefficients acquired at 50 frames/sec. Several different visual processing methods were examined in order to gain insight into which lip deformations were the most beneficial for speech recognition. All visual features resulted from the projection of the lip outline represented as a sequence of control points onto a sub-space spanned by a recognition basis. The three recognition bases investigated were the affine basis and two bases obtained from principal components analysis.

6.2.3   Affine Basis

The first recognition experiments were conducted with the visual features expressed in terms of affine deformations of the lip template. Artificial Gaussian noise was added to the audio signal, post-segmentation, until a desired signal-to-noise ratio (SNR) was reached. In these experiments SNRs of 0 dB (noise power = signal power), -3 dB and -6 dB were used. Addition of the noise post-segmentation facilitated the audio-only recognition problem as word endpoint detection was accomplished without the need for complex thresholding algorithms. However, the HMM recognisers discussed in the next section attempted recognition without this knowledge of the word boundaries.

The exemplar templates were created from the clean audio and visual data, so no explicit knowledge or assumptions were made about the noise. Although as a practical matter, the normalisation done during audio feature extraction partially compensated for some of the effects of the spectrally flat Gaussian noise and would probably not have a similar compensatory effect on other types of noise.

Error rates using the affine basis for audio signals at various SNRs with different sound-to-vision weightings are shown in figure 6.13. Several points are evident from this graph. The first is the remarkable robustness to noise of the audio-only recogniser. This is most likely the result of the way the noise was added (post-segmentation) and the normalisation step done during feature extraction. Secondly, the audio-only recogniser performs better than the visual-only recogniser particularly at high signal-to-noise ratios (6% error rate versus 52%). This merely reflects the higher information content in audio data with respect to speech recognition. Thirdly, incorporation of the vision information improved performance at all noise levels — providing the most benefit at high levels of degradation — a key finding of this research. It is this increase in recognition performance due to the incorporation of visual information, or *incremental vision rate*, that is a true measure of the added benefit of lip reading.

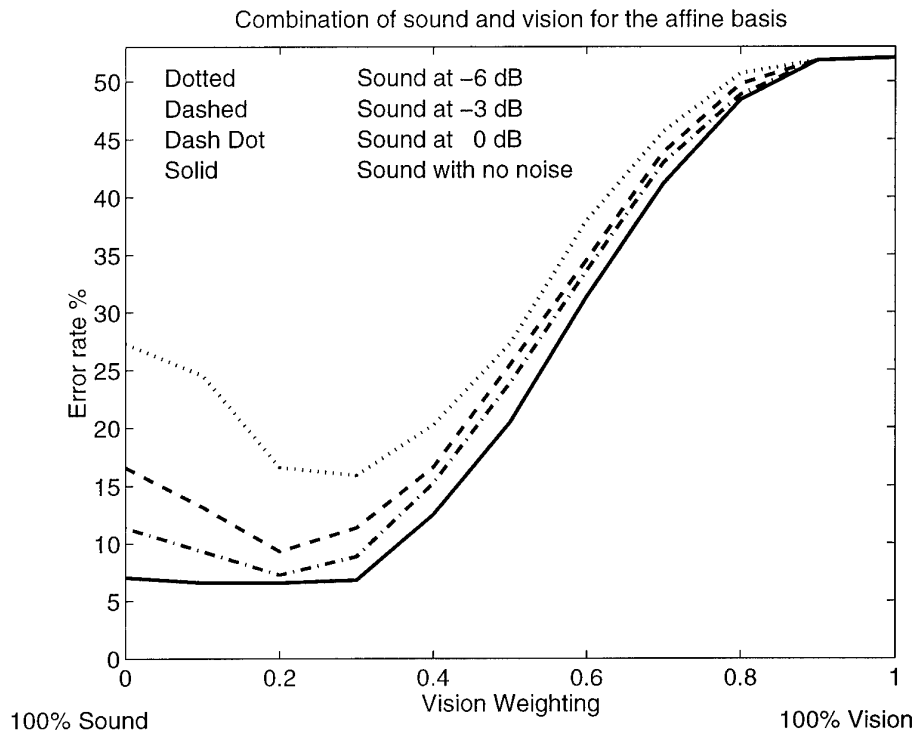A chart summarising the recognition performance, including the incremental vision rate,

Figure 6.13: *Frontal View: Error rate variation on the test set as sound-to-vision weighting is varied. With a clean audio signal, vision is only marginally beneficial — improving the error rate by 0.5%. However as the signal becomes more noisy, the contribution of vision is noticeably improved with a reduction in error rate from 12% to 7% for the 0 dB signal and from 17% to 9% for -3 dB. With the audio quality further degraded to -6 dB, the error rate drops from 27% to 16%.*

for the audio-only, visual-only, and audio-visual systems with the acoustic data at various
SNRs is shown in table 6.5. It is evident that further degradation of the audio signal

Error Rates using Affine Basis					
	Audio test	Visual test	A-V test	Inc Vision Rate	Error Rate Reduction
clean	6.0%	52.0%	5.5%	0.5%	8.3%
0 dB SNR	12.0%	52.0%	7.0%	5.0%	41.7%
-3 dB SNR	16.6%	52.0%	9.3%	7.3%	44.0%
-6 dB SNR	27.0%	52.0%	16.0%	11.0%	40.1%

Table 6.5: *Recognition performance of the 3 DTW recognisers at various acoustic noise levels. Incorporation of the visual information improves performance at all noise levels — providing the most benefit the further the degradation.*

would lead to a decrease in audio-only recognition performance and a further increase in
the incremental vision rate, but the -3 dB signal-to-noise ratio will be used as a standard
comparison point.

6.2.4   Principal Component Bases

Similar experiments to those conducted using the affine basis were accomplished using the
bases derived from principal components analysis — the *PCA basis* and the *PCA no X basis*
(section 6.1.2). The PCA basis comprised the first six principal axes of lip motion, while
the PCA no X basis used the first six principal axes after removing the global horizontal
motion. Only the first six components were used as they accounted for over 99% of the
lip motion. The lip contour expressed as a set of control points was projected onto the
sub-spaces spanned by these bases and the resultant vectors used as visual feature vectors.
Representative lip deformations for these bases are shown in figures 6.4 and 6.6.

The best error rates achieved using visual data expressed in each of the PCA bases
are shown in table 6.6 on sound at -3 dB SNR. (The results from the recognition experiments
with the affine basis are included as well for comparison.) All three bases provide a
similar increase in recognition performance with the error rate of the acoustic-only recogniser
(16.6%) being nearly twice that of the audio-visual recognisers (9.3%–9.8%). These
results demonstrate that there is useful recognition information contained in the lip outline,
contrary to Bregler et al. [17, 19] who found the outline of the lip too coarse for accurate
recognition. Furthermore, the comparable performance of the affine basis with respect to the
derived bases suggests the possibility of developing a multi-speaker or speaker-independent
recognition system with the visual features represented as affine transformations of the lip

Best Error Rates for each basis (Sound at -3 dB)							
Basis	Audio		Visual		A-V		Inc Vision Rate	Error Rate Reduction
	train	test	train	test	train	test		
affine	13.9%	16.6%	44%	52%	8.2%	9.3%	7.3%	44.0%
PCA	13.9%	16.6%	42%	51%	9.6%	9.3%	7.3%	44.0%
PCA no X	13.9%	16.6%	41%	49%	9.6%	9.8%	6.8%	41.0%

Table 6.6: *Best error rates for the three recognition bases on sound at -3 dB SNR. The error rate of the acoustic-only recogniser (16.6%) is nearly twice that of the audio-visual recognisers (9.3%–9.8%) — that is, incorporation of the visual shape parameters results in a 44% reduction in error rate. Further, all three bases provide a similar increase in recognition performance. This is encouraging as the geometrically derived affine basis presents an opportunity for speaker-independent recognition, while the PCA bases are particular to a given speaker.*

template.

6.2.5   Evaluating Visual Shape Components

Having determined the utility of lip information, the potential recognition benefit from only a single vision component was examined. It was hoped that a coherent picture would result yielding the lip movements most beneficial for speech recognition. Table 6.7 shows the recognition performance achieved using only a single vision component from each of the bases. Error rates are shown for the components used individually and in concert with the acoustic features. The tests, which were conducted on speech at a SNR of -3 dB, present several messages.

One can see that the first component of the PCA and PCA no X bases, which represents movement of the lower jaw, contributes substantially to recognition performance. This movement is expressed in the affine basis as a combination of Y scale and Y translation, and it was expected that the Y scale component would afford the better results as the Y translation component can be corrupted by head movements. Surprisingly, it was the Y translation component that yielded the higher incremental vision rate. This may merely reflect the fact that the opening and closing of the mouth can be thought of as a change in the displacement of the lip centroid (Y translation). PCA component 4 and PCA no X component 3, which represent the degree of curling of the lip corners, similarly contribute to recognition performance. This movement represents a non-affine deformation and hence there is not a corresponding affine deformation against which to compare it; however, the Y shear (rotation) deformation exhibits a degree of lip curling, potentially accounting for its surprisingly good performance. The last movements contributing appreciably to recognition performance are the affine X scale and PCA component 5. While it would be premature

Best Error Rates using only a single vision component					
Basis component	Vision only training	test	A-V (Sound at -3 dB) training	test	Inc Vision Rate
Full affine	44%	52%	8.2%	9.3%	7.3%
X Trans	93%	93%	14%	17%	0.0%
Y Trans	76%	81%	13%	14%	3.0%
X Scale	59%	63%	9%	12%	5.0%
Y Scale	75%	79%	14%	16%	0.7%
Y Shear	77%	86%	14%	13%	3.2%
X Shear	86%	90%	14%	17%	0.0%

Full PCA	42%	51%	9.6%	9.3%	7.3%
1	70%	74%	11%	12%	4.6%
2	91%	91%	14%	17%	0.0%
3	82%	88%	14%	17%	0.0%
4	70%	75%	9%	11%	5.2%
5	73%	82%	12%	12%	4.6%
6	89%	94%	14%	17%	0.0%

PCA no X	41%	49%	9.6%	9.8%	6.8%
1	69%	73%	11%	12%	4.6%
2	82%	89%	14%	17%	0.0%
3	67%	71%	9%	12%	5.0%
4	76%	84%	14%	17%	0.0%
5	84%	82%	14%	15%	1.3%
6	82%	90%	14%	17%	0.0%

Table 6.7: *Results of recognition performance using only one vision component from each of the 3 bases. Recognition using sound alone at -3 dB was 13.9% for the training set and 16.6% for the test set. Full affine, Full PCA, and PCA no X refer to overall recognition performance using all six components of each basis. The lip deformations represented by PCA components 1,4,5, PCA no X components 1,3, and affine components Y Trans, X Scale, and Y Shear appear to contribute the most to recognition performance implying that the recognition information of the lip outline can be expressed with just a few shape parameters.*

to draw any strong conclusions concerning their potential use in recognition, since they account for less than 2% of the variance present in lip movements, we saw in section 6.1.4 that the X scale component (horizontal elongation of the mouth) was repeatable across multiple repetitions of the same word, and it is possible that such subtle lip movements may possess important recognition information.

6.2.6 Conclusions

Speaker-dependent, isolated-word recognition experiments have demonstrated that shape parameters obtained from accurately tracked lip contours provide a rich source of information for audio-visual speech recognition. The incorporation of this visual data into acoustic recognisers enables robust speech recognition in the presence of high levels of interfering noise. In addition, the comparable performance of the affine basis with respect to the PCA bases suggests the possibility of developing a multi-speaker or speaker-independent recognition system with the visual features represented as affine transformations of the lip template. Further, an interesting finding of the recognition experiments using individual shape components was that, although typically 6–8 components are necessary for accurate tracking, the recognition information tends to be concentrated in only three shape parameters. These deformations correspond to the degree of mouth opening, the amount of curling of the lip corners, and the horizontal elongation of the mouth, and can be roughly equated to 'ah', 'ee', and (vaguely) 'oh'.

The experiments reported thus far have used dynamic time warping as the recognition algorithm; however, given the state of the art in speech analysis [111], it is natural to try Hidden Markov Model recognition. The next section details experiments using the HMM-based recognisers.

6.3 Hidden Markov Model Recognition

The purpose of the HMM recognition experiments were three-fold. First, it was hoped to confirm the findings of the DTW experiments, that is, that real-time lip tracking can provide valuable information to audio-visual speech recognisers, using state-of-the-art speech recognition methods. Secondly, the HMM-based recognisers serve as a bridge to the more complex task of multi-speaker, continuous speech recognition, which could naturally follow from this research. Lastly, the recognition experiments provide insight into some of the practical problems facing commercial audio-visual recognition systems, where unknown, varying levels of noise may be present, and where the word boundaries are not known *a*

priori and must be determined by the recogniser. Surprisingly, since much of the audio-visual recognition work uses artificially added noise (including the work presented here), as opposed to noisy speech gathered in a natural setting, like a rowdy pub or a noisy office, it is common for researchers to segment the speech by hand or use the clean acoustic data for segmentation, and hence the endpoint detection problem is often not addressed.

Towards these ends, the HMM recognisers were developed and operated as connected-word recognisers, looking for "sequences" of words in the following form: *silence-word-silence*. This enabled investigation of the difficult endpoint detection problem, while still maintaining consistency with current state-of-the art continuous-speech recognisers, where continuous speech is recognised as a connected sequence of phonemes or tri-phones. In this research, whole words were used as the recognition unit instead of phonemes or tri-phones which is the more common practice. This was done primarily for practical reasons (a lack of facilities for, and expertise in, segmenting and labelling audio-visual data). However, the use of whole words had a practical advantage in that feature sequences were about 30 frames long, instead of the 7 or 8 that would have resulted with phoneme/viseme decomposition. This permitted easy capture of distinctive coarticulation effects. From a recognition standpoint, the theory and implementation of connected-word and connected-phoneme recognition are identical, and thus, the system is directly extensible to continuous-speech applications.

The recognition experiments conducted using the HMM recognisers were intended to investigate the potential problem areas of connected-speech audio-visual recognition. Thus, it was known beforehand that direct comparison of the HMM and DTW systems would be difficult. Specifically, although both recognition systems were intended to illustrate the benefit of incorporating lip shape information into acoustic speech recognisers, the HMM system was forced to determine word boundaries (end points), whereas the DTW was provided that information.

6.3.1 Ten Word Database

As a starting point, audio-only (AO), visual-only (VO), and combined audio-visual (AV) recognition experiments were conducted on a 10 word database consisting of people's names (table 6.8). Thirteen repetitions of each word were recorded with 9 used for training and 4 for testing, making the training set 90 words and the test set 40 words. It was known that the small size of this database might preclude making any broad conclusions about vision's contribution to speech recognition; nevertheless, the database serves as a vehicle for efficiently testing the HMM recognition platform and identifying areas requiring further

Ten Word Database									
Alexis	Barney	Charlie	David	Edward	Frederick	Gerald	Harriet	Ian	John

Table 6.8: *A 10 word database consisting of people's names used for preliminary tests into the recognition potential of lip contours in speech recognition applications. Thirteen repetitions of each word were recorded.*

study.

6.3.2 Affine Basis

The HMM model set consisted of 11 models — one for each of the 10 words in the database and an eleventh model corresponding to the no-speech condition (silence). Six-state, left-to-right HMMs, which allowed transitions only to successive states or back to the same state, were used for each of the word models. A single-state HMM was used for the silence model. Observation densities for each of the states in the models were represented by multivariate Gaussians with diagonal covariance matrices. Trials were conducted using multinomial distributions (mixtures of Gaussians) and a single Gaussian with a full covariance matrix, although it was determined that the increased number of free parameters resulted in the over learning of the training data. Eight mel-scale filter bank coefficients were used as audio features, while the visual feature vectors corresponded to affine deformations of the outer lip contour. Composite audio-visual feature vectors were obtained by concatenating the 8 audio and 6 visual features.

Recognition results for the audio-only, visual-only, and audio-visual systems on clean and noisy (10 dB SNR) speech using the affine basis are shown in table 6.9. In clean conditions,

Error Rates using Affine Basis						
	Audio		Visual		A-V		Inc Vision
	training	test	training	test	training	test	Rate
clean	0.0%	0.0%	2.2%	15.0%	0.0%	0.0%	—
10 dB SNR		40.0%	2.2%	15.0%		23.1%	16.9%

Table 6.9: *Overall results for the audio-only, visual-only, and audio-visual HMM recognisers on clean and noisy speech using the affine basis. The models were trained using clean audio and visual data and thus assumed no explicit or implicit knowledge of the noise. The error-free performance on the clean acoustic data can be attributed to the small size of the database. The 23.1% error rate attained on the noisy audio-visual data, despite being an improvement on the acoustic-only recognition (40.0%), was higher than was expected and warranted further investigation.*

the error-free performance of the audio-only and audio-visual recognisers was not surprising given the small size of the database; however, the relatively poor recognition performance

of the audio-only recogniser (40.0% error rate) and the combined audio-visual recogniser (23.1% error rate) on the noisy speech data was higher than expected. In particular, the error rate of the combined audio-visual recogniser, although better than that attained using only the audio data, was worse than that achieved using the visual data alone (15.0%).

A possible explanation of these results can be found in the Viterbi state alignments for correctly and incorrectly identified utterances. Figures 6.14 and 6.15 show the Viterbi state
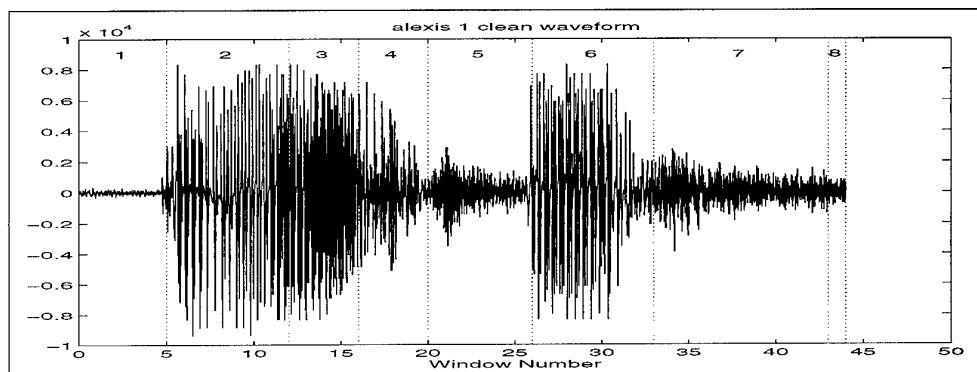
Figure 6.14: *Viterbi state alignment for audio waveform of "Alexis" recognised using* **audio-only** *models. The first and eighth states (dotted lines) represent the transitions from silence to the word and from the word back to silence. Note that the state transition boundaries correspond to acoustically identifiable phases of the utterance.*

alignment on the speech waveform and mel-scale filter banks using the audio-only trained models for a correctly identified "Alexis". One can see that the recogniser has successfully partitioned the utterance into 'silence' (state 1), 'A-le' (states 2-4), 'x' (state 5), 'i' (state 6), 's' (state 7), and 'silence' (state 8). Similar plots of the affine data using the visual-only trained models are shown in figure 6.16. It is difficult to identify any specific transition regions in the vision signal other than noting that that the phonemes / K / S / IH / S / representing 'xis' have been grouped into a single state (6) corresponding to frames 21-41. This was initially unexpected, but upon further inspection deemed entirely appropriate. After positioning the lips for the 'x', the 'is' is produced primarily by modulating the air flow with subtle movements of the tongue. The bottom lip moves slightly downward in the transition from 'x' to 'i' (frames 26-30), but essentially the lips remain stationary during the 'xis' articulation. As should have been anticipated, the Baum-Welch training has correctly grouped the lip movements corresponding to the four phonemes into a single long-duration state. From a word recognition viewpoint, this should represent a distinctive characteristic of the "Alexis" word model. However, potentially strong visual coarticulation effects, such as observed here, may pose a problem for sub-word (phoneme/viseme) recognition. A recent
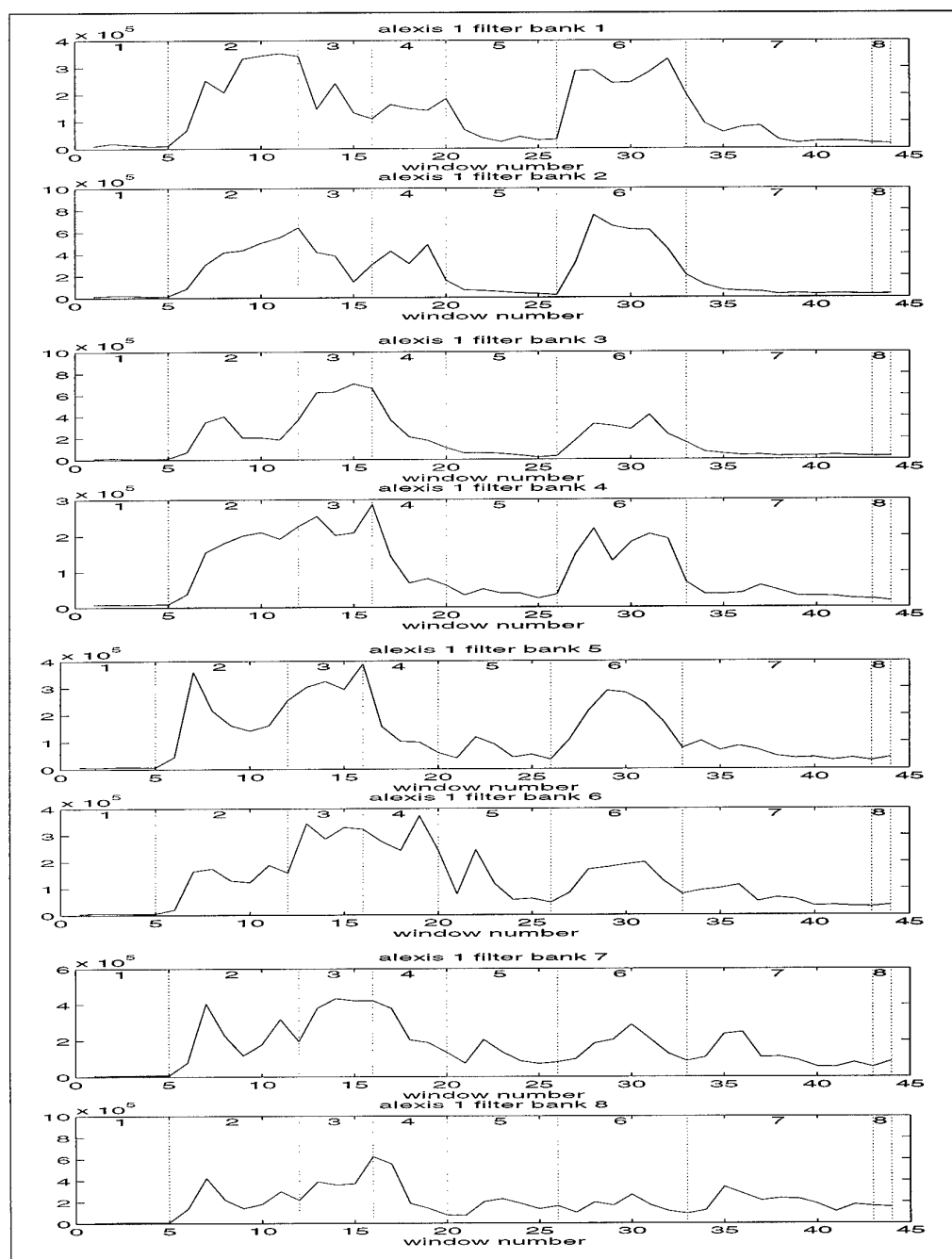
Figure 6.15: *Viterbi state alignment of the mel-scale filter banks for clean speech "Alexis" recognised using* **audio-only** *models. The first and eighth states (dotted lines) represent the transitions from silence to the word and from the word back to silence. Note how the sequence has been partitioned into regions of relatively constant frequency magnitude. For instance, the low energy 'x' (frames 20-26) is readily identifiable.*
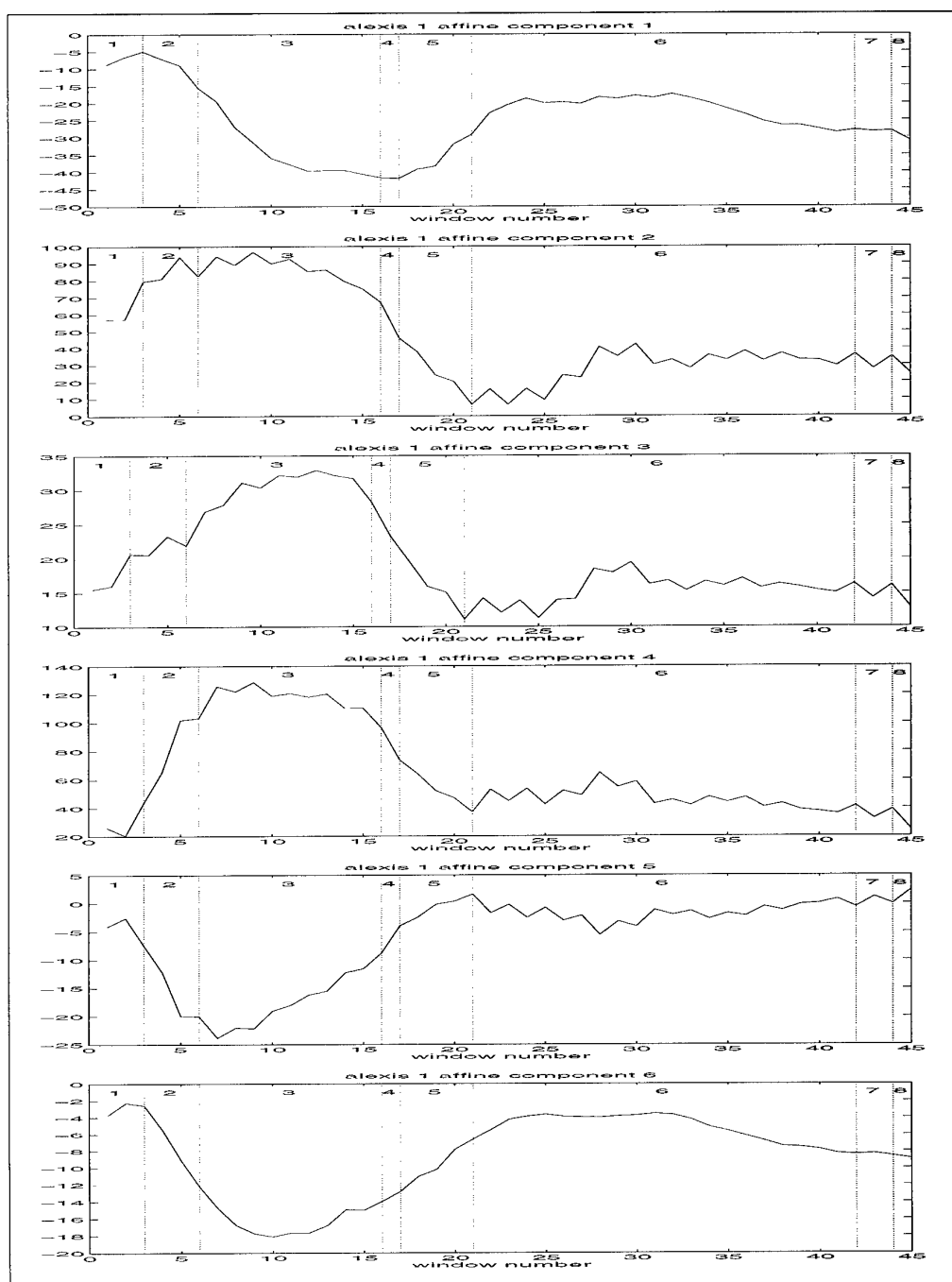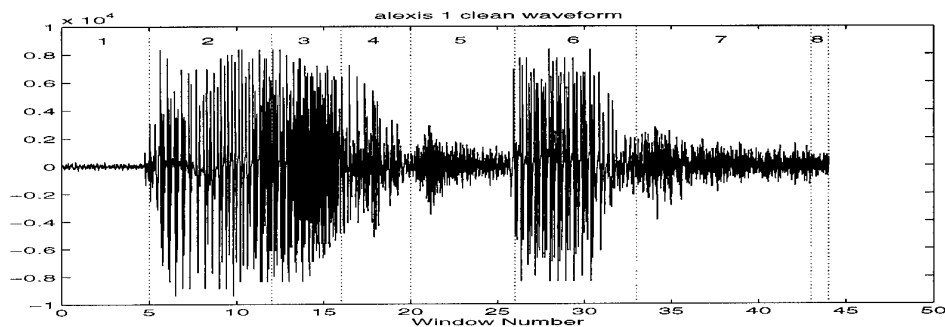
Figure 6.16: *Viterbi state alignment of the Affine components for "Alexis" recognised using visual-only models. The first and eighth states (dotted lines) represent the transitions from silence to the word and from the word back to silence. Note that the phonemes / K / S / IH / S / representing 'xis' have been grouped into a single state (6) corresponding to frames 21-41. This was initially unexpected, but after further analysis deemed entirely appropriate, as after the initial forming of the 'x', the lips remain relatively stationary throughout the 'is'.*

audio-visual integration strategy proposed by Tomlinson et al. [133] using *cross-product* HMMs, which allows for asynchrony between the audio and visual components, may be able to compensate for these effects.

Returning to the audio-only recognition results of table 6.9, the high (40%) error-rate on the 10 dB SNR speech using only the audio-only models was not unexpected as the HMMs were trained with clean speech and then tested on noisy speech; that is, there was a mismatch between the training and testing conditions. In effect, the word models were repetitively trained with filter-bank coefficients clustered about a mean corresponding to clean speech, which resulted in models with corresponding mean vectors μ and small variances Σ. They were then presented a noisy signal with a different mean and a large variance to recognise. Not surprisingly, the result is poor recognition performance. One method for dealing with these mismatched conditions is to adjust the word models at recognition time by incorporating noise models appropriate to the perceived level of noise [52, 53].

A more disturbing result was the relatively poor performance (23.1% error rate) of the combined audio-visual recognition system, especially in light of the 15.0% error rate achieved using only the visual data. Examination of the Viterbi state alignment obtained using the learnt audio-only, visual-only, and audio-visual models overlayed on the audio and visual signals provides one explanation. This is illustrated in figures 6.17 and 6.18.
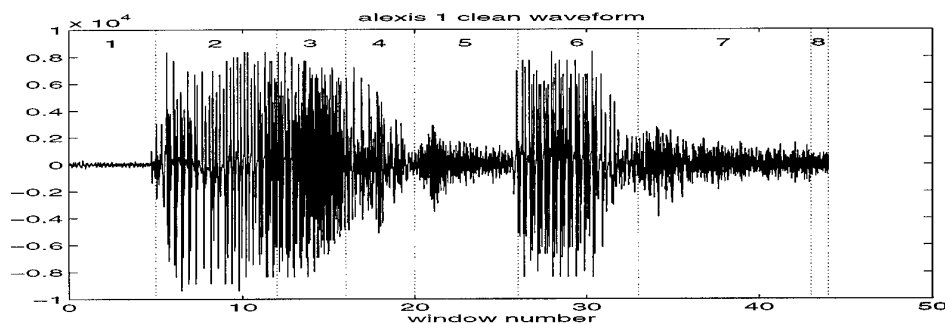
In figure 6.17, note how the state alignment is identical for the audio-only and audio-visual trained models. This suggests that the audio-visual models may be tuned more to the acoustic data than the visual data, which results in a Viterbi alignment that closely follows the acoustic transitions in the speech. When both the acoustic and visual signals are noise-free, the higher linguistic information content of the acoustic channel should naturally result in models which reflect this reality; however, when the acoustic channel is potentially noisy and the visual signal essentially noise-free, training on clean audio-visual data and then testing on noisy audio and clean visual data results in less than optimal performance. Figure 6.18 illustrates this using clean and noisy (10 dB SNR) renditions of "Edward". Edward was correctly identified using the AO, VO, and AV models when no artificial noise was added. However, when the audio signal was corrupted by additive noise, it was incorrectly recognised as "David" with state alignment shown in figure 6.18c. Note how the unmodelled acoustic noise has rendered the audio-only recogniser ineffective at even locating the start of the word (frame 5). When the clean visual signals are appended to the noisy acoustic features, the combined audio-visual recogniser is still unable to provide correct boundary identification (figure 6.18d), despite having been correctly recognised as "Edward" using only the visual data. Essentially, the poor match between the noisy audio data and the

(a) Audio-Only Models
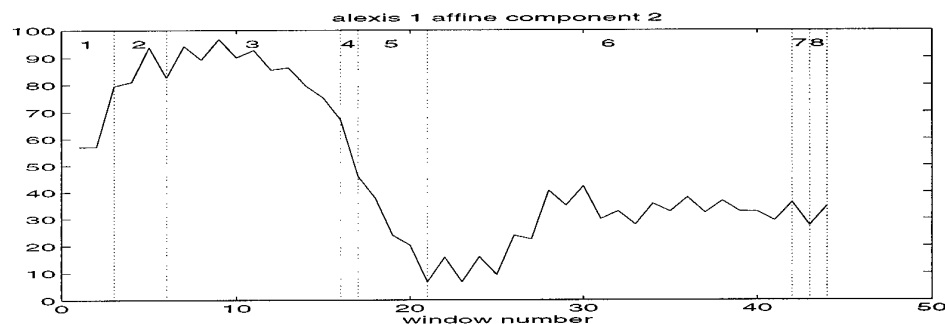
(b) Visual-Only Models

(c) Audio-Visual Models

Figure 6.17: *Viterbi state alignment for clean speech waveforms of Alexis recognised using audio-only, visual-only, and combined audio-visual data. Note how the audio-only (a) and combined audio-visual (c) segmentations are identical. This suggests that the audio-visual models may be tuned more to the acoustic data than the visual data. When the acoustic signal is noise-free, such learning is appropriate; however, when noisy audio-visual speech is to be recognised using HMMs trained on clean speech, the performance is less than optimal (see figure 6.18).*
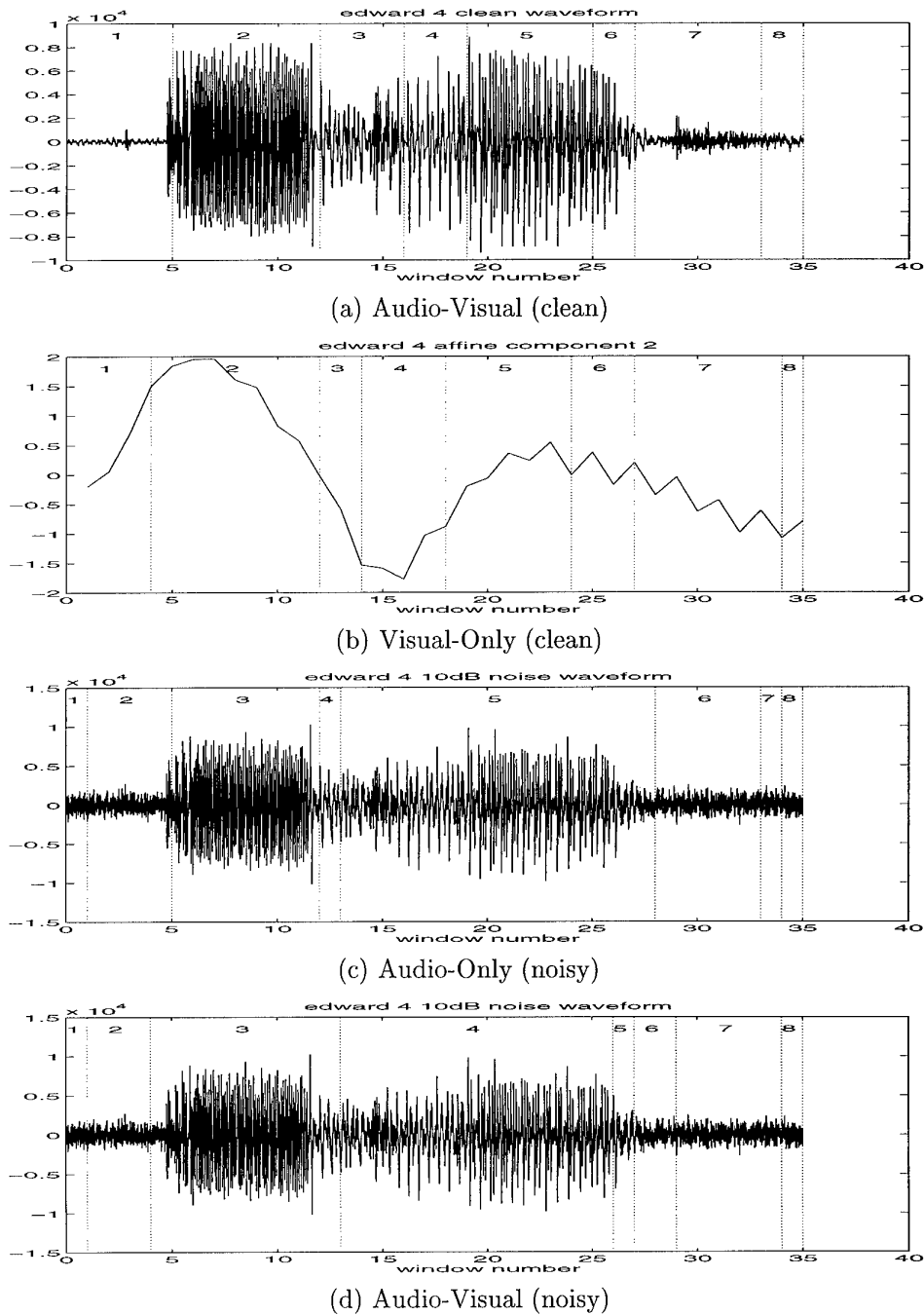
(a) Audio-Visual (clean)

(b) Visual-Only (clean)

(c) Audio-Only (noisy)

(d) Audio-Visual (noisy)

Figure 6.18: *Unmodelled noise in the acoustic signal can result in misidentification of the audio-visual boundaries despite the presence of the visual signal. Shown are the Viterbi state alignments for correctly recognised clean speech, (a) and (b), and incorrectly recognised noisy speech, (c) and (d), of "Edward". The audio-only and audio-visual models were trained using noise-free acoustic conditions. The presence of unmodelled acoustic noise results in misidentification of "Edward" using the acoustic models (c). Note how even the start of the word (frame 5) is incorrectly detected as frame 1. Compare with (a) where the start of the speech is correctly identified. Incorporation of the visual signal is insufficient to overcome the effects of the unmodelled noise (d), for once again the start of the word is inaccurately identified and the result is incorrect audio-visual recognition.*

learnt audio models has pulled down the visual-only recognition.

A simple partial fix to this problem was found by observing that many of the recognition errors of the AO and AV recognisers were due to their inability to correctly identify the start of the words (state 1, frames 0-5 in figure 6.18). This was due to the recognisers being unable to recognise the (noisy) period of "silence" prior to the start of the word. To compensate for these segmentation errors, a new model, representing "noisy silence", which was trained on the background noise, was introduced into the model set. The result was a reduction in the number of segmentation errors, and an increase in the overall audio-only and audio-visual recognition rates.

The benefit provided by the addition of the noisy-silence model can be seen in figure 6.19 where the sequences of figure 6.18 are recognised using the increased model set. Although recognition of "Edward" using the audio-only models was still incorrect (figure 6.19c), the start of the word was correctly identified. The improved boundary detection provided by the noisy-silence model enables the incorporation of the visual signal to provide the sought after increase in recognition performance. This is illustrated in figure 6.19d where the audio-visual models result in correct recognition and an appropriate state segmentation.

6.3.3 Noise Compensation

While the use of the noisy-silence model reduced the number of segmentation errors due to misidentification of the word boundaries and resulted in improved audio-visual recognition, it represents only the first step in tackling the more fundamental problem of how to appropriately deal with the presence of an unknown level of noise in the acoustic channel. The most straightforward approach, which does not assume any explicit knowledge of the noise, is to train the combined audio-visual models with speech utterances at various signal-to-noise ratios to represent the uncertain nature of the noise level in the audio signal. This should lead to relatively large variances in the speech feature vectors compared to those of the clean visual signal. For some applications, there may not be any *a priori* knowledge of the noise type or level, and hence this strategy may more closely approximate the real-world speech recognition problem; however, this approach ultimately begs the question "How many different types of noise (artificial Gaussian, crosstalk, ringing telephones, ...) and noise levels (clean, 10 dB, 0 dB, ...) should be included in the training set?" The answer to that question depends primarily on the expected environment of the given application and, in any case, requires a large amount of training data.
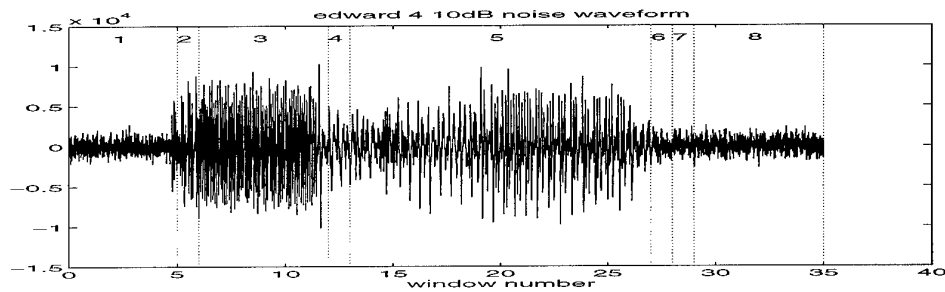
Recently, several strategies for dealing with the effects of degraded speech have been proposed [21, 74, 53]. These noise compensation approaches vary in the degree of assumed
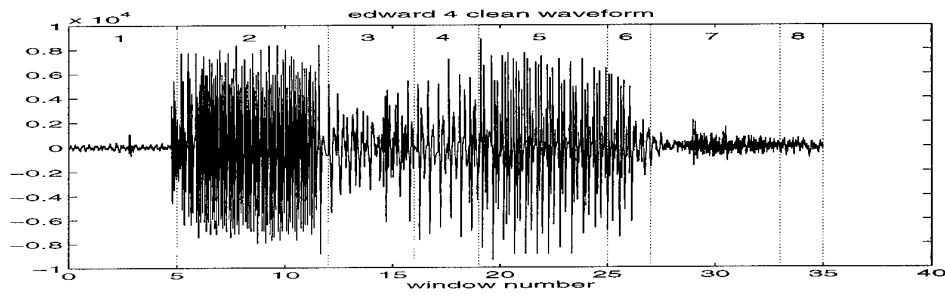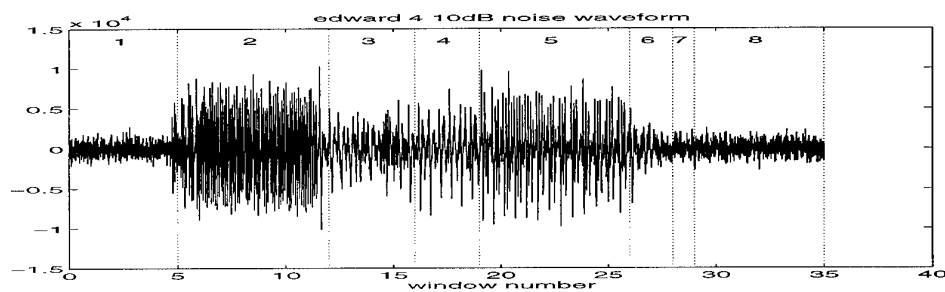
(a) Audio-Visual (clean)
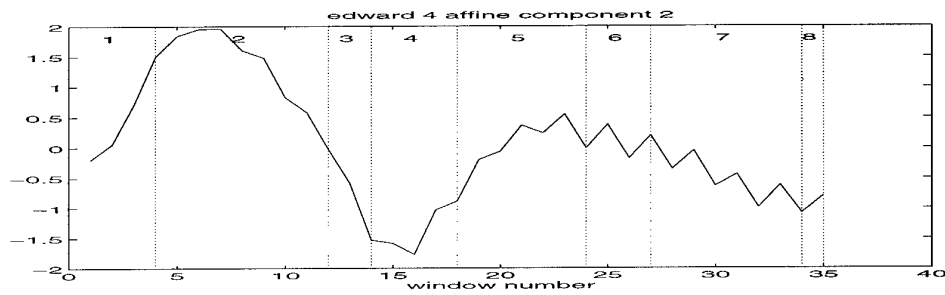
(b) Visual-Only (clean)

(c) Audio-Only (noisy silence)

(d) Audio-Visual (noisy silence)

Figure 6.19: *Use of a "noisy-silence" model to limit boundary detection errors enables the incorporation of the visual data to provide correct audio-visual recognition. "Edward" was incorrectly recognised using the audio-only models (c), although the start of the word was correctly identified (frame 5). Compare this with the correctly segmented clean signal (a). Incorporation of the visual signal results in the desired effect as the noisy "Edward" (d) is correctly recognised with an appropriate state segmentation.*

knowledge of the noise level and type, and the stage (model building or recognition) at which this knowledge is utilised. The three primary compensation strategies are briefly explained below and more detailed reviews can be found in [103, 137]. The first approach uses filtering or noise masking or cancellation techniques on the noisy speech prior to feature extraction [80, 21] . Essentially noise tracking is used to determine the level of the noise and then an estimate of the speech spectra is obtained by subtracting the estimate of the noise spectra from that of the noisy speech. The second approach attempts to use novel speech features which are robust in the presence of noise [65, 137]. One technique, based on the premise that recognition features which explicitly encode the temporal dynamics of speech are inherently more robust to noise, utilises 2-dimensional (cepstral-time) feature matrices to encode the short-time variations of the speech cepstral coefficients [136, 137]. The third main noise compensation strategy dynamically adapts the speech models during recognition based on the estimated level of noise. This parallel model combination approach [135, 52, 53] uses separately trained speech and noise models which are married at recognition time under the assumption that the speech and noise are additive in the linear power domain.

Another strategy for dealing with the complicating effects of noise-degraded speech, while not a noise compensation technique *per se*, is to use a noise tracker to estimate the level of the noise and then use audio-visual models for recognition that were trained at the appropriate noise level. However, it may be impractical to require training of the HMMs at every possible noise level for all of the likely noise sources. Further, acoustic noise in natural environments, such as an office or laboratory, is likely to be non-stationary and occur in short bursts, such as a co-worker initiating a conversation or a telephone that occasionally rings. Accordingly, the recogniser would need to be able to dynamically switch recognition models with the perceived changes in noise level. Despite these practical problems, training directly on the noisy speech does have several distinct theoretical advantages. First, it enables the learning algorithm to generalise over the linguistically relevant audio-visual information, resulting in optimal information extraction at a given noise level. Secondly, it eliminates the difficulties associated with a mismatch in the training and testing conditions which may lead to unexpected or less than optimal results, as was illustrated earlier. Lastly, acoustic recognition performance attained from training and testing in like conditions represents an upper bound for acoustic recognition performance, that might be attainable with a highly tuned, noise-compensated, acoustic speech recogniser. Thus, any gains provided by incorporation of the visual information into a noise-trained acoustic recogniser, represents the *minimum* contribution that can be attributed to the visual data.

A final recognition experiment was conducted on the 10 dB SNR speech where the

audio-only and audio-visual speech were trained using the noisy speech. The results of this experiment, as well as the other two recognition experiments where no noise compensation was used (table 6.9) and where a noisy-silence model was used (figure 6.19) are shown in table 6.10. Although the type of noise compensation method employed resulted in a fairly substantial change in the final audio-visual recognition error rates (0.0%–23.1% for this database) and the corresponding incremental vision rates (2.5%–16.9%), in all cases, the incorporation of visual shape information resulted in improved recognition performance.

Error Rates using various noise compensation techniques						
noise compensation	Audio training	test	Visual training	test	A-V training	test	Inc Vision Rate
none		40.0%	2.2%	15.0%		23.1%	16.9%
noisy silence		23.1%	2.2%	15.0%		10.0%	13.1%
known noise level	0.0%	2.5%	2.2%	15.0%	0.0%	0.0%	2.5%

Table 6.10: *Error Rates using various noise compensation techniques. The type of noise compensation method employed can result in a fairly substantial change in the final audio-visual recognition error rate and the corresponding incremental vision rate. In the first experiment, no noise compensation was employed. Although the visual information substantially decreases the error rate from 40.0% down to 23.1%, the lack of noise compensation results in an unacceptably high error rate. At the opposite end of the spectrum, when the noise level is assumed known, the visual information reduces the error rate by 2.5% down to error-free performance. An intermediate approach, which utilises models at several different noise levels to represent the period of "silence" separating words, but no noise compensation for the word models themselves, results in slightly more than a halving of the error rate (23.1% to 10.0%). Despite differences in the absolute error rates, the incorporation of the visual information has resulted in improved recognition accuracy in all situations.*

The acoustic recognition of speech in adverse environments and the development of novel noise compensation strategies remains an active area of research, which is receiving increased attention [96, 53, 137]. As demonstrated above, precise quantification of the contribution afforded by the visual channel can be difficult as error rates are affected by the noise compensation strategy used. Thus, since training and testing on speech at the same noise level represents an upper bound on the acoustic recognition performance, subsequent recognition experiments were conducted under these conditions. Recognition results on the 10-word names database using matched training and testing conditions at various acoustic noise levels are shown in table 6.11. Although the resultant audio-visual error rates are better than one might expect to achieve using a more natural noise setting, the added benefit afforded by the visual signal is clear.

Error Rates using known noise level							
	Audio		Visual		A-V		Inc Vision
	training	test	training	test	training	test	Rate
10 dB SNR	0.0%	2.5%	2.2%	15.0%	0.0%	0.0%	2.5%
-3 dB SNR	0.0%	5.0%	2.2%	15.0%	0.0%	2.5%	2.5%
-6 dB SNR	2.2%	7.5%	2.2%	15.0%	0.0%	5.0%	2.5%

Table 6.11: *Matched training and testing conditions results in low error rates for severely degraded speech. Despite the high acoustic recognition rates, incorporation of the visual information provides improved performance.*

6.3.4 Forty Word Database

The recognition experiments using the 10-word database validated the HMM recognition platform and illustrated the consequences which could be attributed to the use (or lack of use) of noise compensation techniques on the Viterbi state alignment. However, it still remains to confirm the findings of the DTW recognition experiments on a larger database using the more advanced, connected-word, HMM recognisers.

Error Rates using Affine Basis								
	Audio		Visual		A-V		Inc Vision	Error Rate
	train	test	train	test	train	test	Rate	Reduction
clean	0.0%	1.3%	13.0%	33.7%	0.0%	1.3%	0.0%	0.0%
-3 dB SNR	8.4%	11.7%	13.0%	33.7%	3.2%	8.7%	3.0%	25.6%
-6 dB SNR	16.2%	24.6%	13.0%	33.7%	5.0%	10.0%	14.6%	59.4%

Table 6.12: *Recognition results for the HMM recognisers on clean and noisy speech using the affine basis. Incorporation of the visual information enables robust recognition of degraded speech. Further, the utilisation of parameter uncertainty inherent in HMM recognition provides increased recognition accuracy when compared to the recognition using DTW (cf. table 6.5).*

Error Rates using PCA Basis								
	Audio		Visual		A-V		Inc Vision	Error Rate
	train	test	train	test	train	test	Rate	Reduction
clean	0.0%	1.3%	10.5%	25.0%	0.0%	1.3%	0.0%	0.0%
-3 dB SNR	8.4%	11.7%	10.5%	25.0%	3.2%	7.9%	3.8%	32.5%
-6 dB SNR	16.2%	24.6%	10.5%	25.0%	4.6%	10.8%	13.8%	56.1%

Table 6.13: *Recognition results for the HMM recognisers on clean and noisy speech using the PCA basis. Again, incorporation of the visual information enables robust recognition of degraded speech. The error rates using the PCA recognition basis are very similar to those attained using the affine basis (table 6.12).*

Recognition experiments were conducted on the 40-word database (table 6.4) used in

the DTW recognition experiments (section 6.2). Fourteen repetitions of each word were used for training, and 6 for testing, resulting in a training set of 560 words and a test set of 240 words. Error rates attained using the affine and PCA recognition bases are shown in tables 6.12 and 6.13, respectively. The results re-affirm the DTW findings that visual shape parameters can supplement acoustic speech recognisers, enabling robust recognition in degraded acoustic environments. Additionally, the error rates achieved using the HMM recognisers were better, across the board, when compared to results attained using the DTW recognisers (tables 6.5 and 6.6). The most substantial gain was evident in the visual-only recognition (52% vs. 34% for the affine basis and 51% vs. 25% for the PCA basis). The principal advantage that the HMM recogniser has over its DTW counterpart is that the variability, and hence uncertainty, in the visual components is modelled and then utilised in the pattern matching. The result is a recogniser that is better able to capture the linguistically informative traits of the visual features.

6.4 Conclusions

Analysis of natural, continuous speech acquired using a dynamic contour lip tracker suggests that half of the lip movement present in speech is due to the opening of the mouth. Recognition experiments confirm that this motion does indeed serve as a rich source of information for audio-visual speech recognition. Further experiments suggest that the width of the lip and the degree of curling of the lip corners also provide useful linguistic information. These three deformations roughly correspond to 'ah', 'ee', and 'oh'.

Despite doubts expressed by other researchers [17, 19], recognition experiments using dynamic time warping and Hidden Markov Model-based recognisers demonstrate that shape parameters obtained from accurately tracked lip contours can be used to make speech recognition robust to high levels of interfering noise. In noisy acoustic conditions, error rate reductions up to 44% are realised. Further, although the experiments have been conducted on only a single speaker, the comparative performance of the affine basis with the more specialised PCA bases suggests the possibility of developing a multi-speaker or speaker-independent recognition system with the visual features represented as affine transformations of the lip template.

7

Colour Lip Tracking

In chapter 4 it was shown that the dynamic contour tracking framework permits accurate, real-time tracking of the outer lip contour. The recognition experiments of chapter 6 verified that shape information extracted from the outer lip contour provides a rich source of information for audio-visual speech recognition. Further tracking experiments on unadorned lips showed that more powerful feature detection methods could be employed to attain accurate tracking of unadorned lips. In particular, it was shown that a data-driven approach based on statistical models of the grey-level intensity profiles, together with a Mahalanobis distance measure, could be used to successfully identify the lip boundary in grey-level images (section 4.5). Lip tracking was stable and accurate, although the analysis suggested that tracking performance was bounded by the limited information content in grey-level images of the mouth region. Here alternate feature detection methods are presented which take advantage of the differing pigmentation of the skin and lips. Discriminant analysis and Bayesian classification on colour images of the face are used to identify features that correspond to the boundary between the lips and the surrounding skin and mouth. Utilisation of these feature detectors results in accurate, outer lip contour tracking which is robust to variations in lighting as experienced in an uncontrolled office environment. Further, modelling of the inner mouth colour intensities enables tracking of both the inner and outer lip contours.

7.1   Facial Colour

The perceived colour of an object is a linear combination of the colour of the light reflected from its surface (specular reflection) and the colour of the light reflected from its body (diffuse reflection). In the case of facial skin, the colour is primarily determined by the amount of melanin in the skin and the blood beneath the epidermal layer. As such, it is usually found in a restricted range of hues [119]. In fact, the variation in skin colour between members of a given race is small enough that it has been effectively used to identify people in a wide variety of images [50, 79]. Colour imagery has also proven effective for locating faces and initialising face trackers [45, 127]. Typically the colour information is used to provide coarse identification of regions of the face and then more sophisticated image processing is used to fine tune the positioning of facial features. Here a different approach is taken. Pattern classification techniques are applied to colour images of the face to accurately pinpoint the lip boundaries, overcoming the limitations inherent in greyscale images.

7.2   Hue Discrimination

Perceptually, when we humans think of the *colour* of objects, we typically think in terms of hue, saturation, and value (or intensity). The hue of an object is a polar angle corresponding to the nearest "pure" colour, where red, green, and blue are traditionally given values $0°$, $120°$, and $240°$, respectively. The perceptual notion of hue can be thought of as the dominant wavelength present in the observed light. Saturation refers to how tinted a colour is, that is, the ratio of pure colour to white light. For example, red is highly saturated, pink relatively unsaturated, and grey completely unsaturated. Finally, value corresponds to the achromatic notion of intensity and represents the largest of the red, green, and blue channels. Since skin colour typically takes on only a restricted range of hues, representing colour images of the face in HSV (hue, saturation, value) space has a certain appeal. Indeed, Fleck et al. [50] use hue-based discrimination for identifying skin coloured regions within images. Despite its intuitive appeal, problems exist with the HSV representation. Specifically, at low values of saturation, the hue can vary widely for seemingly similarly coloured regions. Further, when the saturation falls to zero (white or gray), the hue is undefined. Finally, there is an additional problem in that being angular, hue is discontinuous at $360°$ ($0°$). For these reasons, when using the HSV representation, it is not uncommon for researchers to use hand-set heuristics to deal with these limitations. For instance, Fleck et al. [50] adjust the range of acceptable hue values in their skin classifier as a (presumably non-linear) function of saturation. Similarly, Vogt [139] uses an empirically determined non-linear 2D LUT (look

up table) as a function of hue and saturation to segment lip regions within colour images of the face.

Although there may be specific problems with the HSV representation, the successful use of colour to provide coarse localisation of people and faces in images suggests that there is sufficient information in colour images of the face to identify the lip-skin boundary. Rather than devising methods for overcoming the inherent limitations of the HSV representation, it was hoped that a more automatic method for identifying the lip boundary in colour images could be found. Fortunately, as was illustrated in earlier chapters, because of the shape and motion modelling inherent in the Kalman filter framework, it is not necessary to achieve error-free classification or segmentation in order to achieve accurate tracking, rather it is only necessary to find features that mostly correspond to the boundary. It was hoped that these shape and motion models could replace the hand-tuning needed by the other researchers.

7.3 Colour Image Feature Detection

Surprisingly, despite our intuitive feel for a colour "edge", the concept is not well defined. Edge detection in grey-level images has been much researched [91, 28, 41]; however, little formal work has been done on colour edge detection. Nevatia [102] proposed a method where separate 1D edge detectors were run over each of the three colour channels. Colour edges were then identified as areas producing responses in multiple channels.

An alternate approach is to formulate the identification of the boundary between objects, or within an object, as a pattern classification problem [36]. When the object to be tracked is significantly different in colour from its surround, such as identifying oranges for robotic harvesting [109] or distinguishing cabbages from soil for autonomous robot guidance [140], fast, simple, discrimination methods can be employed. For example, Wildenberg [140] used the fact that cabbages always appear green, while soil tends to have a reddish brown tint in nearly all lighting conditions, to formulate a simple red versus green test to separate cabbages from their surrounding soil.

Accurate demarcation of the lip-skin boundary is a significantly harder problem, principally because the lips and skin, although different, are very similar in colour. Further, there is rarely an abrupt change in colour from facial skin to lips; rather, the transition is often more gradual and it is quite common for individuals to possess a ragged outer lip boundary as opposed to a nice smooth one. The challenge is how best to make efficient use of the additional discriminating potential of colour.

7.4 Feature Identification via Bayesian Classification

Identification of the boundary between the lips and their surround can be formulated as a pattern classification problem. Specifically, it is desired to classify colour pixels as to whether it is more likely that they came from the lips or the neighbouring regions (eg. surrounding facial skin, inner mouth region). In a Bayesian framework, a colour pixel, $\mathbf{x} = (r, g, b)^T$, represented as a vector of red, green, and blue intensities, is classified as belonging to the lips if its *a posteriori* probability $P(\text{lips}|\mathbf{x})$ is greater than the corresponding *a posteriori* probability for the surrounding skin, $P(\text{skin}|\mathbf{x})$, and vice versa if

$$P(\text{skin}|\mathbf{x}) > P(\text{lips}|\mathbf{x}).$$

If the class-conditional probability densities $p(\mathbf{x}|\omega_j)$, where ω_j represents the class from which the pixel originated (eg. $\omega_1 = \text{skin}$, $\omega_2 = \text{lips}$), are known, or can be learnt from training images, then Bayes Rule

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})}, \tag{7.1}$$

where $p(\mathbf{x})$ is given by

$$p(\mathbf{x}) = \sum_{j=1}^{2} p(\mathbf{x}|\omega_j)P(\omega_j)$$

can be used to compute the *a posteriori* probabilities.

Standard parametric or non-parametric techniques can be used to learn the underlying class-conditional densities $p(\mathbf{x}|\omega_j)$. However, one must bear in mind that the *a posteriori* probabilities $P(\omega_j|\mathbf{x})$ are evaluated for each pixel, along each search line, at each time step. Thus, in order for this approach to be usable in practical (real-time) systems, a premium is placed on the time required to discriminate between the classes. Towards this end, Fisher's linear discriminant analysis [46] is used to determine the boundary between the lips and facial skin, that is, identify the outer lip contour.

7.5 Fisher's Linear Discriminant

It was desired to develop a feature detection method that could utilise the full discriminating power of the colour images while providing robustness to changes in illumination without incurring excessive computational overhead. Instead of using hue directly and devising methods for overcoming its inherent limitations, a more general approach was sought. A novel application of Fisher's linear discriminant analysis was used for this purpose.

Rather than assuming a particular form for the underlying class-conditional densities and estimating the distribution parameters from training images, Fisher's linear discriminant analysis is used to project the colour vector data down to a scalar quantity. A major advantage of this approach is that the discriminating power of the colour data can be exploited, while still providing a degree of invariance to changes in illumination. Further, since the learning of the Fisher discriminant is done off-line, there is minimal additional computational load incurred during tracking as scalar feature detectors can be used on the resultant Fisher projection.

Fisher's linear discriminant analysis is traditionally used to reduce the dimensionality of high dimensional data down to a more computationally manageable number. In the simplest case, such as the two class discrimination problem of lips and skin, it can be used to determine an axis, $\mathbf{w}$, onto which vector colour data can be projected which preserves as much of the discriminating capability of the colour information as possible. The result, referred to as the Fisher linear discriminant, maximises the separability of the two classes by maximising the ratio of the between-class scatter to the within-class scatter.

More formally, following [46], let $\mathcal{X}$ denote the set of n rgb-colour pixels $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ in the mouth region, where the subset $\mathcal{X}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}\}$ represents those pixels corresponding to the skin, and similarly the subset $\mathcal{X}_2$ of n_2 pixels $\{\mathbf{x}_{n_1+1}, \mathbf{x}_{n_1+2}, \ldots, \mathbf{x}_{n_1+n_2}\}$ corresponds to the lips. It is desired to find a discriminant axis, $\mathbf{w}$, such that the projection of the samples $\mathbf{x}_i$ onto it, given by

$$y_i = \mathbf{w}^T \mathbf{x}_i \tag{7.2}$$

resulting in $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$, maximises the separability between $\mathcal{Y}_1 = \{y_1, y_2, \ldots, y_{n_1}\}$ and $\mathcal{Y}_2 = \{y_{n_1+1}, y_{n_1+2}, \ldots, y_{n_1+n_2}\}$.

A suitable measure of between-class scatter of the projected points is the difference of their means $|\tilde{m}_1 - \tilde{m}_2|$ where

$$\tilde{m}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in \mathcal{X}_k} \mathbf{w}^T \mathbf{x}.$$

Within-class scatter is found in the traditional manner

$$\tilde{s}_k^2 = \sum_{\mathbf{x} \in \mathcal{X}_k} (\mathbf{w}^T \mathbf{x} - \tilde{m}_k)^2.$$

The total within-class scatter $\tilde{s}_w^2$ is then just the sum of the scatter for each of the classes

$$\tilde{s}_W^2 = \tilde{s}_1^2 + \tilde{s}_2^2.$$

The Fisher linear discriminant is defined as the linear function $\mathbf{w}^T \mathbf{x}$ for which the

criterion-function

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{7.3}$$

is a maximum.

Explicit dependence of J on $\mathbf{w}$ is seen by expanding out (7.3) and defining a few additional terms. Let $\mathbf{m}_k$ denote the mean for each of the classes,

$$\mathbf{m}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in \mathcal{X}_k} \mathbf{x}.$$

Similarly, let S_k denote the within-class scatter matrices for the rgb-colour points, where

$$S_k = \sum_{\mathbf{x} \in \mathcal{X}_k} (\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^T.$$

The total within-class scatter S_W is given by

$$S_W = S_1 + S_2.$$

As above, the between-class scatter of the rgb-colour points is given by the difference of their means

$$S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T.$$

The between-class scatter of the projected points can now be expressed in terms of the scatter matrix S_B,

$$\begin{aligned}
|\tilde{m}_1 - \tilde{m}_2|^2 &= (\mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2)^2 \\
&= \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} \\
&= \mathbf{w}^T S_B \mathbf{w}.
\end{aligned} \tag{7.4}$$

Similarly, the within-class scatter of the projected points can be represented in terms of S_k,

$$\begin{aligned}
\tilde{s}_k^2 &= \sum_{\mathbf{x} \in \mathcal{X}_k} \mathbf{w}^T (\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^T \mathbf{w} \\
&= \mathbf{w}^T S_k \mathbf{w}
\end{aligned} \tag{7.5}$$

from which it follows that

$$\tilde{s}_W^2 = \mathbf{w}^T S_W \mathbf{w}. \tag{7.6}$$

Substituting (7.4) and (7.6) into (7.3) results in the desired dependence

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}. \tag{7.7}$$

The vector $\mathbf{w}$ that maximises the criterion function J, must also satisfy

$$S_B \mathbf{w} = \lambda S_W \mathbf{w},$$

which is a conventional eigenvalue problem. Thus the eigenvector of $S_W^{-1} S_B$ corresponding to the non-zero eigenvalue (S_B is rank 1) solves for $\mathbf{w}$. A simplification results in that $S_B \mathbf{w}$ is in the direction of $(\mathbf{m}_1 - \mathbf{m}_2)$ and hence $\mathbf{w}$ can be found directly from

$$\mathbf{w} = S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{7.8}$$

---

1. **Calculate** the mean pixel values in each class, k=1,2

$$\mathbf{m}_k \;=\; \frac{1}{n_k} \sum_{\mathbf{x} \in \mathcal{X}_k} \mathbf{x}$$

2. **Determine** the within-class scatter, k=1,2

$$S_k \;=\; \sum_{\mathbf{x} \in \mathcal{X}_k} (\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^T$$

3. **Find** the Fisher discriminant vector

$$\mathbf{w} \;=\; S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

where $S_W = S_1 + S_2$.

Figure 7.1: *Learning the Fisher linear discriminant axis. A discriminant vector $\mathbf{w}$ can be learnt from sample colour image data from the two classes, lips ($\mathcal{X}_1$) and facial skin ($\mathcal{X}_2$), which maximises the separability of the two classes by maximising the ratio of the between-class scatter to the within-class scatter.*

The steps for learning the Fisher discriminant axis are shown in figure 7.1 using this simplification. Training images, with the colour pixel data either side of the lip boundary grouped into their respective classes, are used to calculate the discriminant axis by computing class means and within-class scatter using the above algorithm. The ability of the Fisher discriminant to maintain the separability of the classes after projection is illustrated in figure 7.2.

It is also instructive to look at the learnt Fisher axes for different parts of the mouth. For instance, for the middle of the lower lip (figure 7.2)

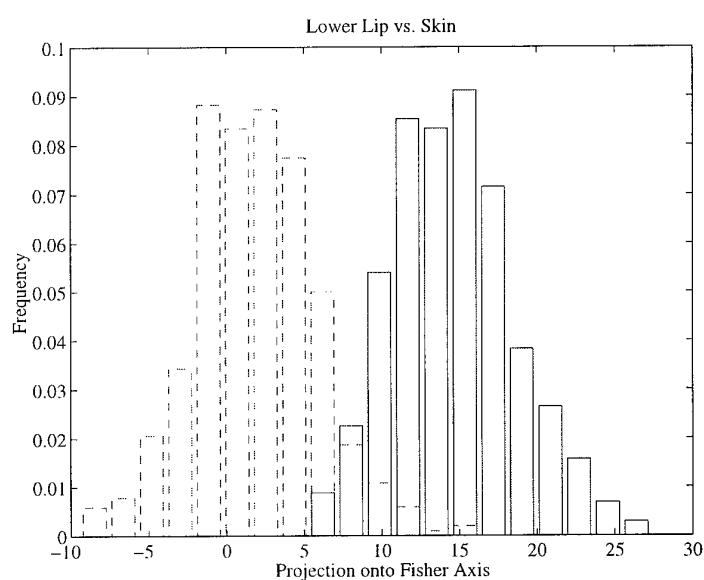$$\mathbf{w} = [-0.0032, 0.6759, -0.7370]^T, \tag{7.9}$$

Figure 7.2: *Discrimination of the lower lip and skin. Colour data from the middle of the lower lip (dashed lines) and surrounding skin (solid bars) are projected onto the learnt Fisher axis. The separability of the data demonstrates the discriminating capacity of the colour information, since the data is drawn from an area of minimal grey-level contrast (cf. figure 7.6).*

which suggests that the classification information for that section of the lip is carried in the green and blue channels. Alternately, for a similar section along the upper lip, the Fisher axis was found to be

$$\mathbf{w} = [-0.3352, 0.8903, -0.3083]^T, \tag{7.10}$$

which implies that in this region all three channels carry information needed for classification. Once again, the projection of the colour data onto the Fisher axis results in well separated classes (figure 7.3).
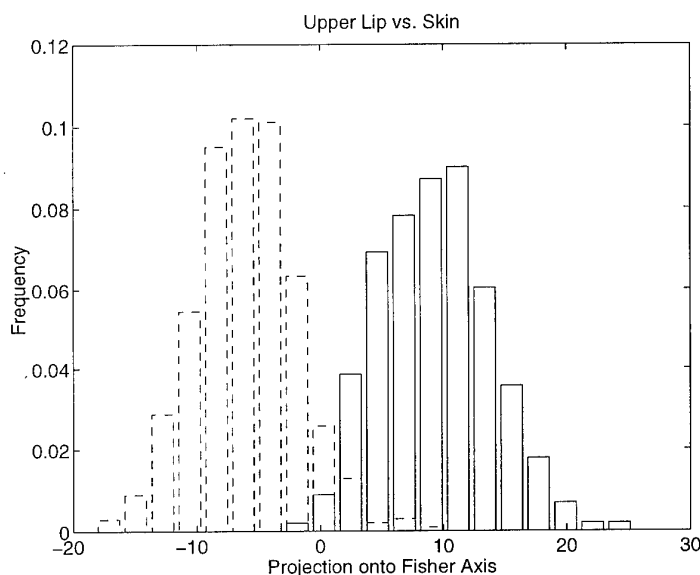

Figure 7.3: *Discrimination of the upper lip and skin. Colour data from the upper lip (dashed lines) and surrounding skin (solid bars) are projected onto the learnt Fisher axis. As with the lower lip, the separability of the data demonstrates the discerning capability of the colour information.*

In contrast to the well separated classes resulting from projection onto the Fisher axes, the two classes are thoroughly intermixed when only intensity information is used. This is illustrated in figure 7.4 where the limited amount of discriminating information in grey-level images is seen by projecting the colour data onto the $r = g = b$ axis, that is

$$\mathbf{w} = [0.3333, 0.3333, 0.3333]^T.$$

In the most general case a separate Fisher axis can be computed for each search line normal; however, in practice, because of the biological consistency of lips and skin, it is only necessary to compute separate discriminants for areas along the mouth that are likely to differ. For instance, if variations in the source lighting are expected and known, such as the primary light source being from the left, then that knowledge and any associated
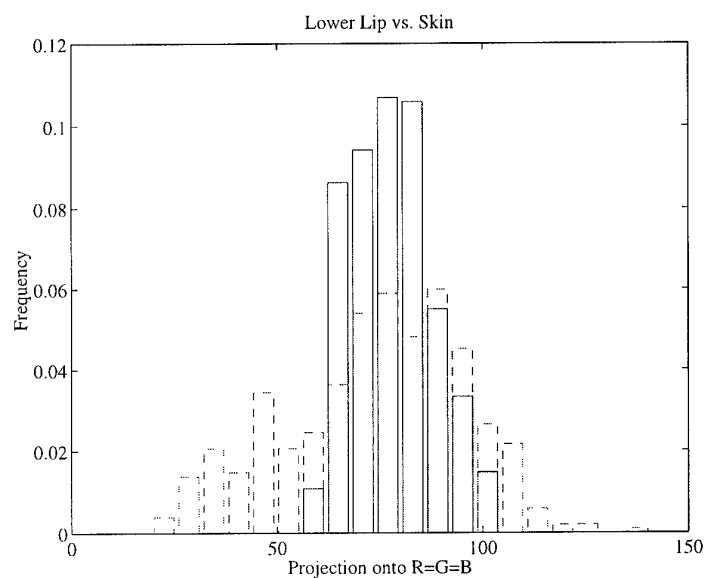
Figure 7.4: *Grey-level intensity data is inadequate for discriminating the lower lip and skin. Colour data from the lower lip (dashed bars) and surrounding skin (solid bars) are projected onto the R=G=B line resulting in their corresponding grey-level value. Note that the two classes are thoroughly intermixed indicating the difficulty in identifying the lip boundary using only grey-level information.*

shadowing can be seamlessly integrated into the discriminant framework by training in like conditions. Similar arguments hold for situations where the principal light source is relatively fixed, such as in a typical office with overhead lighting; however, variations in these cases are more likely associated with changes in environmental conditions such as the relative position of the speaker, camera, and dominant light source. However, in the absence of any specific knowledge of lighting variations, the most straightforward approach is to use a sufficiently large number of training images, where it is hoped that the colour information pertinent to discrimination will be learnt. The result is that data with little discerning information is ignored.

Having computed the Fisher axes, there still remains the question of how best to utilise the resultant scalar projections. One method is to use the statistical template approach described earlier, that is, to learn statistics for the profiles along each of the search lines. However, instead of using the intensities, normalised intensities, or intensity gradients, the Fisher projections can be used. The potential benefit is that the statistical templates might capture additional spatial information along each of the search lines. An alternative method is to continue in the Bayesian classification framework and to make use of the known class-conditional distributions (figures 7.2 and 7.3). If the *a priori* probabilities $P(\omega_i)$ corresponding to the proportion of lip-coloured and skin-coloured pixels are known, or can be estimated from training images, then Bayes Rule (7.1) can be used to compute the *a posteriori* probabilities. Moreover, if it is assumed that the prior probability for a pixel along a given search line is equally likely to be lips as it is to be skin, then the decision boundaries of the density plots obtained from the training images (eg. figure 7.2) can be used directly for classification. In addition, the number of misclassifications can be reduced by post-processing with a median filter. This Bayesian classification approach is illustrated in figure 7.5.

Since hue is often used to locate humans in scenes [50, 79, 127], it is instructive to see if the use of the Fisher discriminant provides any additional benefit over hue-filtered images. Figure 7.6 shows an example image, the colour hue of the image, and the resultant image after projection onto a Fisher axis. Although the hue channel shows high contrast, the localisation of the lip boundary is poor. The Fisher discriminant, on the other hand, results in an image with good contrast and good localisation of the lip boundary.

7.5.1 Environmental Variations

One of the principal advantages of colour images over greyscale images is the added information provided by the three channels as compared to one. Indeed, the previous discussion
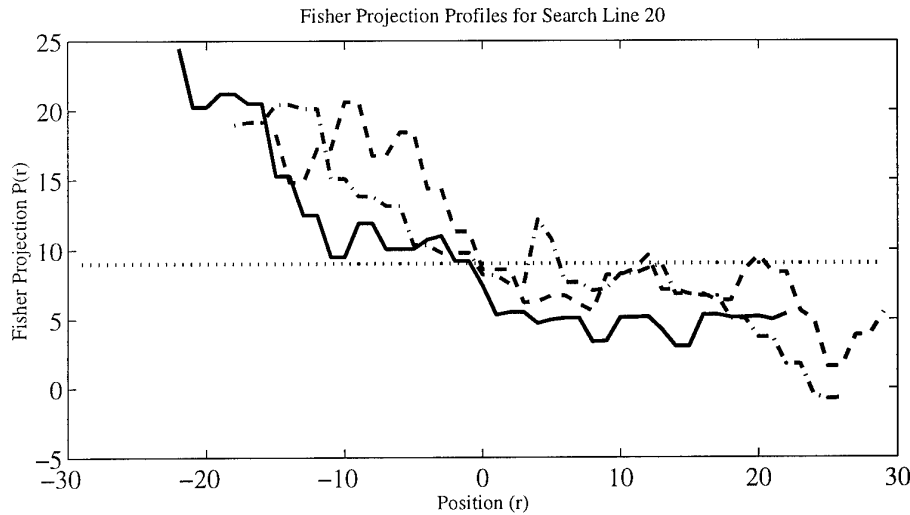
Fisher Projection Profiles for Search Line 20

Figure 7.5: *Classification of the lower lip-skin boundary using Fisher's linear discriminant. Colour pixel intensities along a search line are projected onto a learnt Fisher axis. (Three instances of a search line along the lower lip are shown here to demonstrate the variability over time.) The actual lip boundary corresponds to $r = 0$. Classification is accomplished using the learnt decision boundary (dotted line). The few spurious misclassifications ($r = 4, 12, 20$) are removed by applying a median filter.*

has shown how Fisher's discriminant analysis can be used to capture much of the information in colour images of the face. However, another important aspect of the discriminant analysis approach is what *is not* captured, that is, are uninformative variations ignored?

One of the compelling reasons for operating within a real-time framework is that it imposes a more rigorous standard of acceptability. For instance, having developed a tracker, that is, having created an appropriate shape space, having learnt appropriate motion dynamics, and having learnt Fisher axes for feature detection, it is relatively easy to set up a camera, position the speaker, and track his lips. There will naturally be differences between that particular set up and the conditions present when the motion and image feature models were learnt. Depending on the intended application domain, the variability might be minor or quite substantial. For instance, if one were designing an audio-visual speech recogniser for an outdoor Automated Teller Machine (ATM), the system would have to work on sunny days, cloudy days, rainy days, at night, and so forth. Further, if the system were multi-speaker, in addition to handling inter-speaker visual speech variability, it would have to compensate for the differing heights of the speakers which results in differing viewpoints for the camera. However, even for a speaker-dependent system where the speaker is directed to place his mouth within a particular window, one should expect slight variations in pose
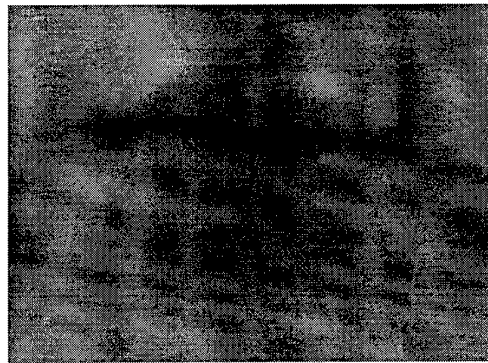
(a) Greyscale Image
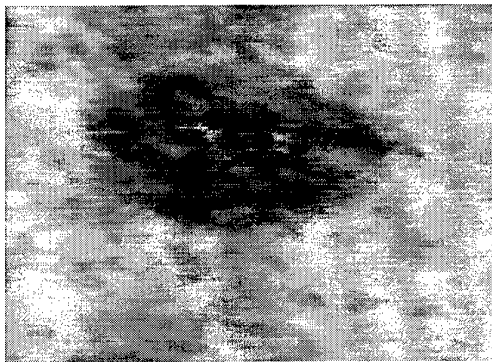
(b) Hue Image
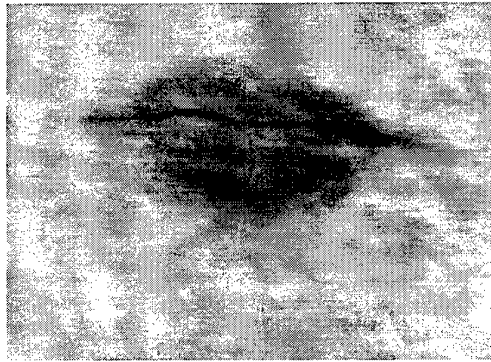
(c) Fisher Discriminant Image

Figure 7.6: *Fisher's discriminant analysis can be used to enhance the contrast between the face and lips. In greyscale images (a) there is little contrast between the face and lips — particularly along the lower lip. The differing hue of the lips and skin can be used to provide additional contrast (b); however, only a coarse identification of the lip-skin boundary is available. Projection onto a Fisher axis (c) enhances the contrast and enables identification of the lip-skin boundary.*

and/or viewpoint angle (say 5°-10°). Audio-visual speech researchers traditionally ignore even these minor variations. For instance, in [17] the database was acquired in a sound-proof booth with controlled lighting to minimise grey-level variations, and presumably to improve tracking performance. It is also not uncommon for researchers to demonstrate tracking on only very short sequences, for instance, a test database of 40 utterances representing less than a minute of speech [62, 64].

It is constructive to look at some of the variation in the training images used in this work, which were acquired in an uncontrolled office environment. An examination of the principal modes of variations, that is, the eigenvectors, of the within-class scatter matrices illustrates the variation in the training images. The eigenvalues and their corresponding eigenvectors for the area of skin along the middle of the upper mouth are

$$
v_1 = \begin{pmatrix} 0.7488 \\ 0.5099 \\ 0.4234 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -0.3567 \\ -0.2284 \\ 0.9059 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 0.5586 \\ -0.8293 \\ 0.0109 \end{pmatrix}
$$

$$
\lambda_1 = 894.7, \quad \lambda_2 = 26.8, \quad \lambda_3 = 13.7.
$$

It is interesting to note that the principal mode of intensity variation (v_1) accounts for over 95% of the variance in the skin data. Closer examination of v_1 reveals that it roughly corresponds to the "pinkness" of the skin. (If the eigenvector were $r = g = b$ then one could say that it corresponded to the variation in grey-level intensity; however, since the eigenvector is tilted towards the red-green sector, one can say that it roughly corresponds to pink variations.) This seems intuitive as the skin represents a roughly homogeneous region, and variations are likely due to changes in the shading of the skin and/or the relative position of the dominant light source (overhead fluorescent lighting in these examples), the camera, and the speaker.

A similar situation exists for the upper lip, where the principal modes of variation and their corresponding eigenvalues given by

$$
v_1 = \begin{pmatrix} 0.7694 \\ 0.5107 \\ 0.3836 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -0.5119 \\ 0.1340 \\ 0.8485 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 0.3820 \\ -0.8492 \\ 0.3645 \end{pmatrix}
$$

$$
\lambda_1 = 792.6, \quad \lambda_2 = 31.9, \quad \lambda_3 = 11.4.
$$

Once again the principal mode of intensity variation (v_1) accounts for over 90% of the variance. As was the case with the skin data, v_1 roughly corresponds to the degree of "pinkness" of the lips. The point to be made is that although there are significant colour intensity variations due to changes in lighting and other environmental variations, the learnt

Fisher axis is not sensitive to these prominent, but un-informative, variations. Indeed, the Fisher axis is nearly orthogonal to the principal mode of within class variation for the lips and skin.

To illustrate the insensitivity of the Fisher technique to environmental variations, additional image sequences were gathered several weeks later with no attempt to replicate the lighting conditions of the original training session. As the intended application is speaker-dependent recognition in a noisy office environment, intensity variations in the images are due to changes in overall lighting, as well as other environmental factors, such as varying camera angles. For instance, in figure 7.7 it can be seen that the two images, (a) and (d), differ both in their overall intensity patterns (in (a) there are significant highlights on the mouth, nose, and chin) and on the degree of self-shadowing (in (a) the shadowing of the upper lip by the nose is more pronounced). However, as such variations are uninformative for identifying the outer lip boundary, projection of the colour intensity data onto the Fisher axis preserves the separability (figures 7.7b and 7.7e). Thus, the strength of the Fisher analysis lies in its ability to capture the most informative aspects of the colour data without being susceptible to variations that have little to do with discrimination, such as changes in illumination.

7.5.2   Outer Contour Tracking

Lastly, when integrated into the dynamic contour framework, use of the Fisher discrimination feature detection method results in robust, accurate, lip tracking which can be accomplished in real-time (50 Hz) on a standard workstation (Silicon Graphics Indy R4400 200 MHz). Accurate tracking was achieved on more than twenty minutes of continuous speech gathered in an office setting without recourse to special lighting. Two example tracked sequences which were recorded on separate days are shown in figures 7.8 and 7.9. In both cases, tracking is accurate through a wide range of lip deformations despite the changes in lighting and camera angle.

7.6   Inner and Outer Lip Contour Tracking

The visual recognition features used thus far contain information only on the positioning of the outer lip contour. As it is known from perceptual studies [131] that human lip-readers rely on information about the presence/absence of the teeth and the tongue inside the mouth, it is natural to try to extract this information from the visual images. In the simplest case, the entire region bounded by the outer lip contour (ie. the whole mouth) can be
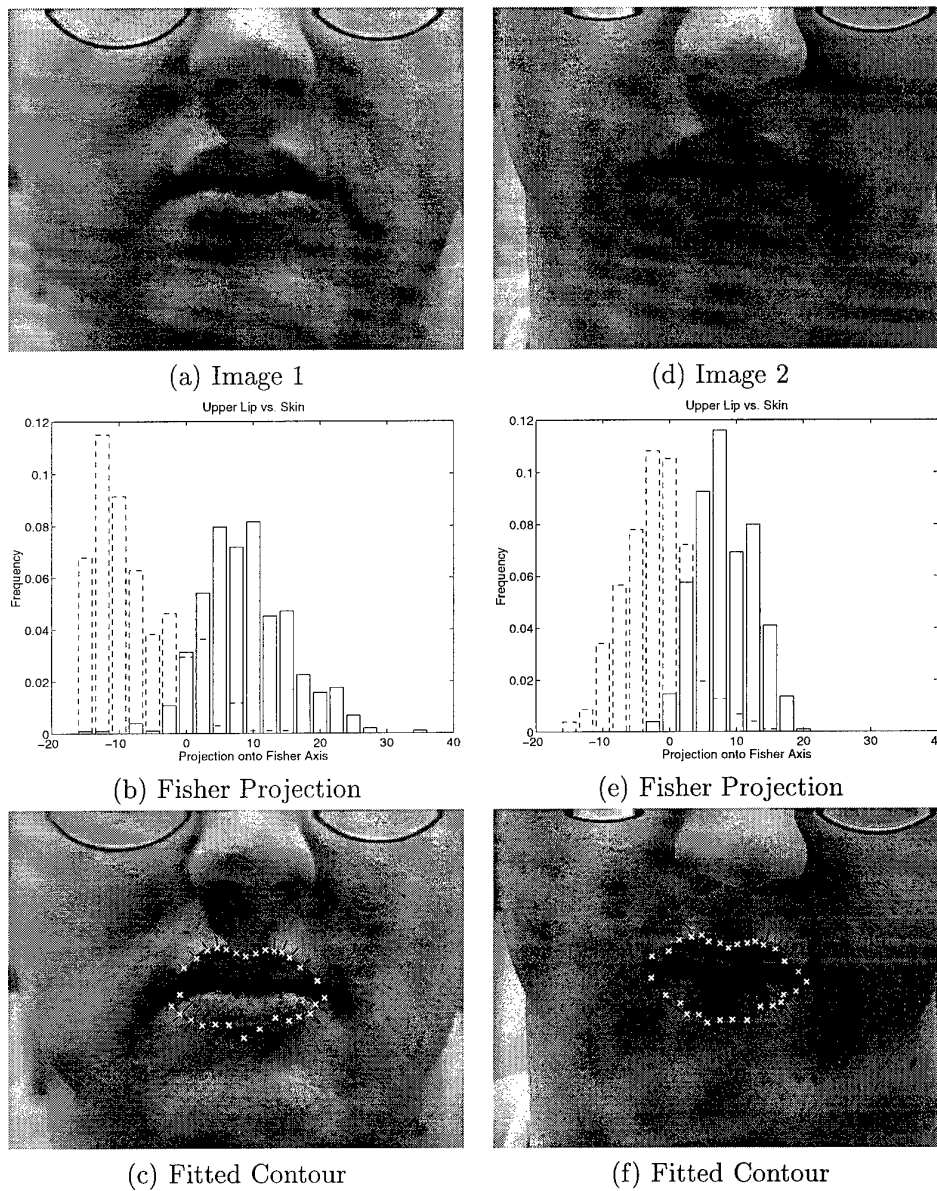
(a) Image 1

(d) Image 2

(b) Fisher Projection

(e) Fisher Projection

(c) Fitted Contour

(f) Fitted Contour

Figure 7.7: *The Fisher axis is robust to environmental variations present in a typical office environment. Since the principal light source is usually fixed in a standard office (eg. overhead fluorescent lighting) intensity variations can result from changes in illumination as well as changes in the relative position and orientation of the speaker and camera. Shown are two images, (a) and (d), acquired on different days and taken from slightly different camera angles. Note that (a) represents a more fronto-parallel view, while in (d) the camera is pointing slightly down at the face (the nostrils are not visible). This results in minor variations in shading as evidenced by the upper lip region beneath the nose, and the highlights along the lower lip (a). However, projection of the upper lip region onto the learnt Fisher axis (7.10) results in well-separated data, (b) and (e). Lastly, (c) and (f) show that the Fisher features correspond nicely to the lip outline as the resultant least-squares fitted contours demonstrate.*
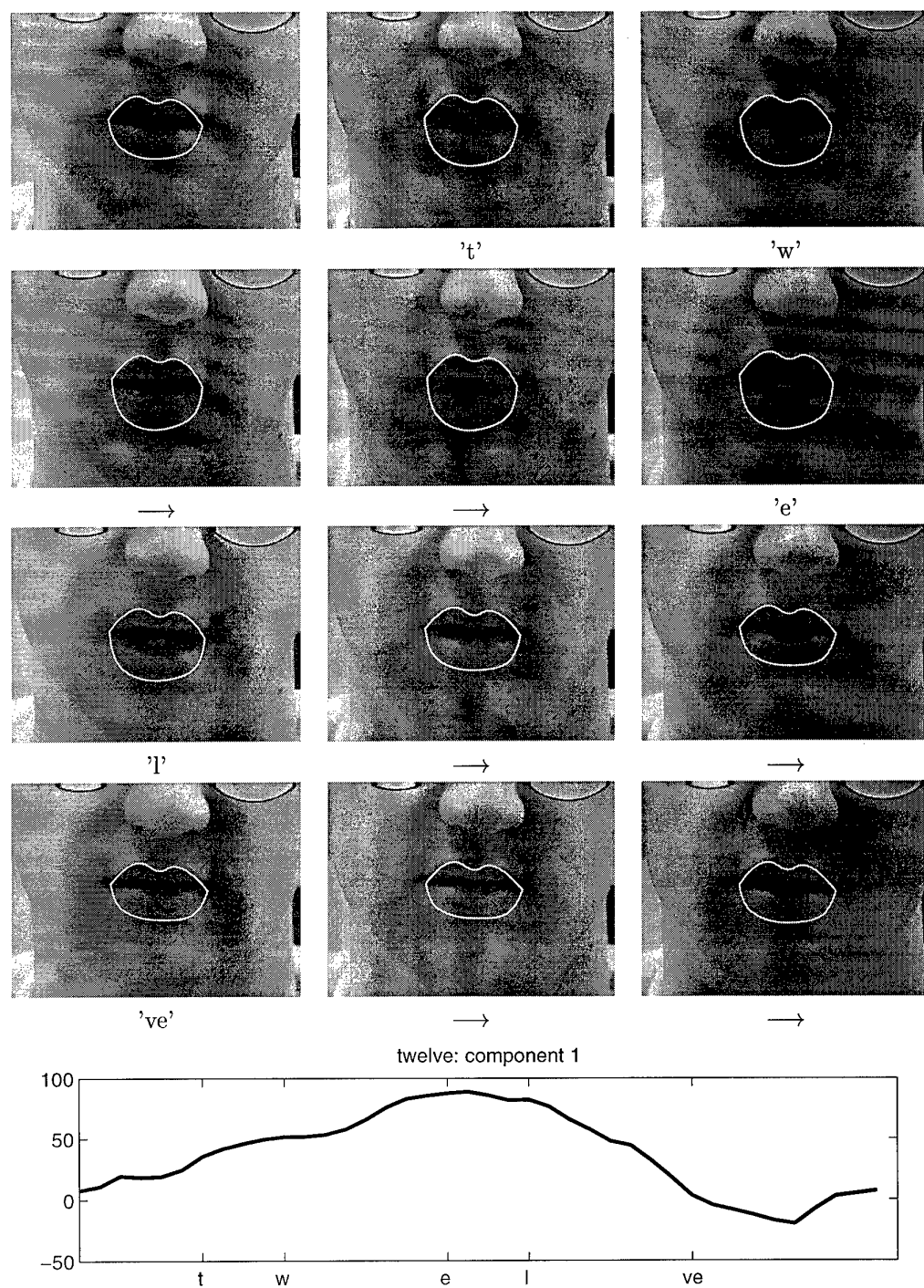
Figure 7.8: *Tracking the word "twelve" on a colour sequence using the Fisher discrimination feature detection method. Snapshots taken approximately every 80 ms. Accurate tracking is attained throughout the sequence, and recognition information, such as the degree of mouth opening (graph), is easily extracted from the tracker.*
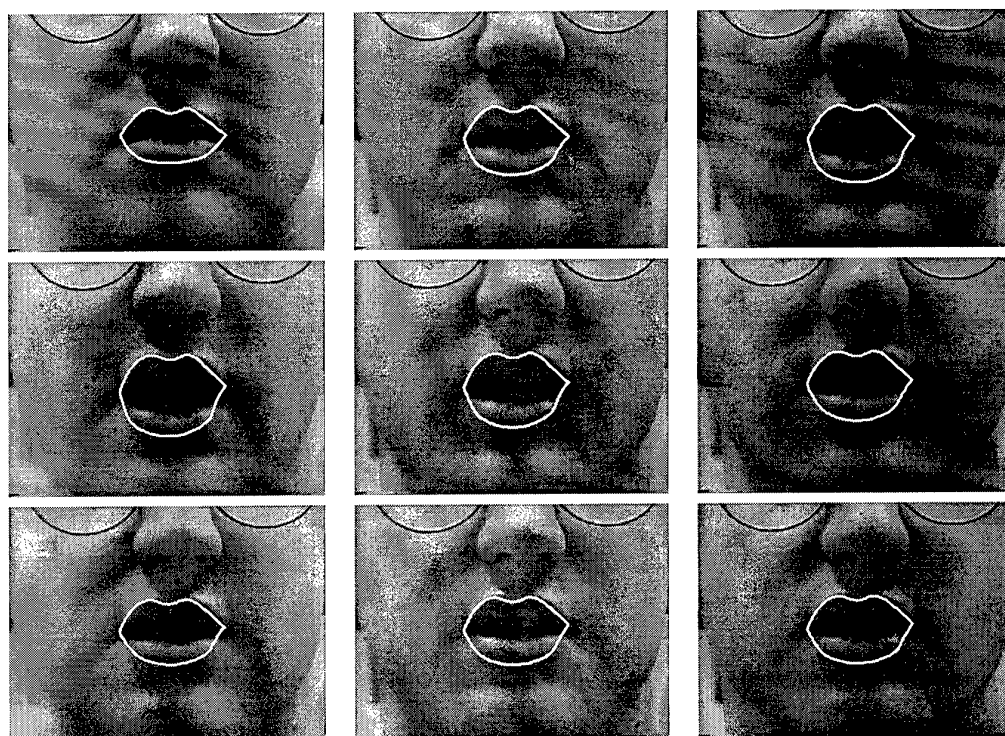
Figure 7.9: *Tracking the word "nineteen" on a colour sequence using the Fisher discrimination feature detection method. Despite the change in camera angle, the tracker accurately follows the lips through the entire sequence with only minor deviations along the upper lip.*

extracted during tracking, and the pixels intensities used as inputs to the audio-visual recognition engine. Others [17, 133] have achieved good audio-visual recognition performance using the pixel intensities directly, although their applications are in controlled lighting conditions. Such an approach may still prove effective in more natural settings, although recognition performance will likely depend on the ability of the classifier to generalise over lighting changes and compensate for tracking errors. A more interesting approach is to attempt to extend the dynamic contour tracking framework to the task of tracking both the inner and outer lip contours. This would permit extraction of the region inside the mouth to more accurately reason about the proportion of teeth visible (both upper and lower), if any, as well as the position of the tongue, if visible. A hybrid system utilising both the contour information, and the mouth region data, would likely provide the best recognition results.

The dynamic contour framework can be extended to the task of tracking two contours, such as both the outer and inner lips, in one of two ways. One approach, after Reynard et al. [117, 78], is to treat the contours as separate objects and explicitly model the dynamical coupling between them. When the coupling between the two objects is causal, such as in their case where the outline of the head was tracked in order to pre-position a mouth-valley tracker, the computational saving of their approach is appealing; however, in a more general setting, such as the bi-directional coupling of the inner and outer lip contours, it is not clear that any computational savings could be achieved. The second approach involves using the constraints provided by the use of restricted shape spaces (section 3.3) to provide the coupling between the inner and outer contours. Since the shape matrix W is learnt from principal component analysis on sample lip shapes, the coupling between the inner and outer contours is directly encoded in the shape matrix (3.5).

7.6.1 Identification of Inner Mouth Region

The principal difficulty in tracking the inner mouth contour is the erratic appearance and disappearance of the teeth. When the teeth are obscured by the lips, there is both an edge and an intensity valley along the inner lip contour [144, 100], but when the teeth are visible, there are numerous edges inside the mouth which serve to distract the tracker (figure 7.10). One method of overcoming this problem is to extend the statistical profile modelling discussed in section 4.5 to handle multi-modal distributions to account for the intermittent presence of the teeth and tongue. An alternate solution is to use the Bayesian classification approach for feature detection as was done using Fisher's linear discriminant.

Modelling of the distribution of colour pixel intensities was facilitated by the observa-
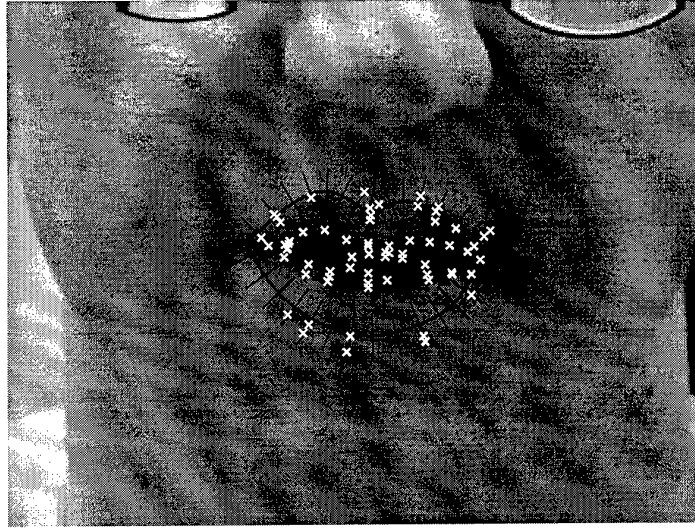
Figure 7.10: *There are numerous distracting edges (white crosses) inside the mouth when the teeth appear.*

tion that there are 3 prominent components inside the mouth, corresponding to a dark region, teeth (upper and/or lower), and tongue (figure 7.11). In any given image, all three components need not be present; however, the entire distribution can be suitably modelled as a mixture of the three components. This suggested a straightforward method for modelling the distribution of colour intensities inside the mouth as a mixture of multi-variate Gaussians,

$$p(\mathbf{x}|\text{inner mouth}) = \sum_{k=1}^{M} c_k \mathcal{N}(\mu_k, \Sigma_k) \tag{7.11}$$

where

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \, e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)} \tag{7.12}$$

and M is the number of mixtures, c_k is the mixture coefficient for the kth mixture and $\mathcal{N}$ is a standard multi-variate Gaussian with mean vector μ and covariance matrix Σ of dimension $d = 3$ for colour vector data.

7.6.2 Expectation Maximisation

The Expectation-Maximisation (EM) algorithm [40, 70] was used to provide maximum-likelihood (ML) estimates of both the mixture weights and the underlying Gaussian parameters. Approximately 50 training images, some with the teeth and tongue present, others with only the teeth present, and still others with only the dark portion of the inner mouth
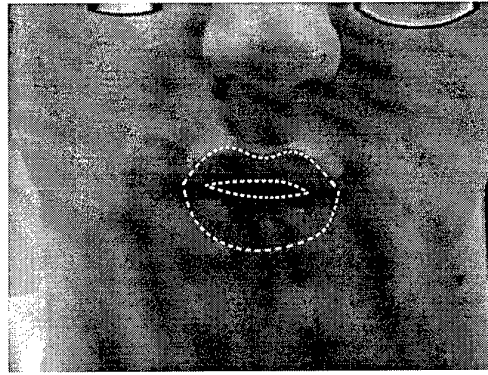
(a) Dark region only
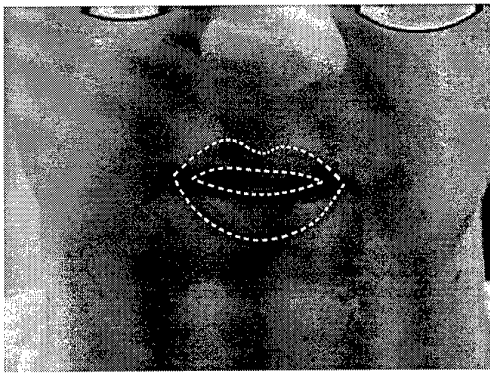
(b) Teeth and dark region
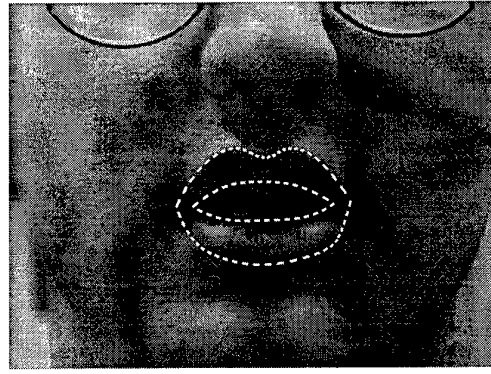
(c) Teeth, Tongue, and dark region

Figure 7.11: *There are 3 prominent components inside the mouth, corresponding to a dark region, teeth (upper and/or lower), and tongue. This suggested modelling the distribution of colour intensities inside the mouth as a mixture of multi-variate Gaussians. Three common situations are illustrated showing various combinations of the components.*

present, were gathered under various lighting conditions similar to those expected to be encountered during tracking. K-means clustering was used to provide initial estimates of the parameters and the EM algorithm was run until convergence. Figure 7.12 gives an overview of the parameter estimation algorithm.

Let $\mathcal{X}$ denote the set of N rgb-colour pixels $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ of the inner mouth extracted from the training images by hand fitting contours to the inner lip outline. The initial cluster centres, $\mathbf{m}_1 \ldots \mathbf{m}_M$, for the K-means clustering were chosen to correspond to each of the three physically significant components inside the mouth, teeth, tongue, and dark region (step 1). In the most general case, the cluster centres can be chosen randomly from $\mathcal{X}$, however, convergence of the EM algorithm is improved when initialised more appropriately. Initial estimates of the mixture weights and Gaussian parameters are computed using the

1. **Select** M initial cluster centres, $\mathbf{m}_1 \ldots \mathbf{m}_M$

2. **Initialise** estimates using K-means clustering, for $k = 1 \ldots M$

$$c_k^{[0]} \;=\; \frac{N_k}{N}$$

$$\mu_k^{[0]} \;=\; \frac{1}{N_k} \sum_{\mathbf{x} \in \mathcal{X}_k} (\mathbf{x} - \mathbf{m}_k)$$

$$\Sigma_k^{[0]} \;=\; \frac{1}{N_k} \sum_{\mathbf{x} \in \mathcal{X}_k} (\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^T$$

3. **Repeat** for $j = 1 \ldots$ Expectation-Maximisation, for $k = 1 \ldots M$

$$c_k^{[j]} \;=\; \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{X}} p^{[j-1]}(k|\mathbf{x})$$

$$\mu_k^{[j]} \;=\; \frac{N}{c_k^{[j]}} \sum_{\mathbf{x} \in \mathcal{X}} p^{[j-1]}(k|\mathbf{x})\mathbf{x}$$

$$\Sigma_k^{[j]} \;=\; \frac{N}{c_k^{[j]}} \sum_{\mathbf{x} \in \mathcal{X}} p^{[j-1]}(k|\mathbf{x})(\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^T$$

until proportional change in $\mu_k < \epsilon$ for each mixture.

Figure 7.12: *Expectation-Maximisation learning of the mixture parameters representing the colour intensity distribution inside the mouth.*

cluster centres (step 2). Pixels are grouped into clusters $\mathcal{X}_1 \ldots \mathcal{X}_M$ based on Euclidean distance to the nearest centre, $\mathbf{m}_k$. The EM algorithm (step 3) is then iterated, where superscript j represents the iteration number, until the proportional change in the means μ_k falls below a set threshold or the maximum number of iterations is reached.

A similar procedure was used to learn the distribution of colour intensities for the upper and lower lips; however, instead of modelling the distributions as mixtures of Gaussians they were modelled as a single 3D Gaussian. Though a single Gaussian may not be suffi-

cient to capture all of the subtle intensity variations within the lip regions, only a coarse representation was needed to adequately discriminate lip coloured pixels from those inside the mouth.

Although the EM algorithm can be computationally expensive, it is run off-line from training images, and thus, during tracking, pixels can be rapidly classified as either inner mouth or surrounding lip using the straightforward Bayesian classifier described earlier (section 7.4). Furthermore, since features are only sought along normals at discretely sampled points along the contour, it is not necessary to classify each pixel in the mouth region — only those along the image normals. The temporal coherence provided by the Kalman filter ensures that only pixels in the mouth region are inspected, permitting use of the two class discrimination solution as opposed to more complicated solutions that might include skin regions, facial hair, scene background, and such. A prototypical example of the classification of the inner mouth region and the resultant fitted contour is shown in figure 7.13.

Although identification of the inner mouth region in the example shown is particularly good, it is important to note that error-free classification, while desired, is certainly not necessary. The feature curve representing the measured position of the lips in the image is the result of a least-squares fit to the image features detected along the normals. This provides a robust measure of the inner lip contour in cases where no features are found along a given normal, or alternately, where misclassification results in incorrectly identified features. Further, the number of misclassifications is minimised by post-processing with a median filter. Lastly, the assimilation of the measured contour with the predicted position in the Kalman filter also smoothes out errors in the measured position.

7.6.3 Tracking

The use of Fisher discriminant axes along the outer lip contour to identify the lip-skin boundary, and colour mixture models to locate the inner contour boundary, when integrated into the dynamic contour tracking framework enables robust tracking (figure 7.14) of continuous speech. Further the tracker is able to run at a near real-time rate of 25 Hz on a standard workstation (Silicon Graphics Indy R4400 200 MHz). An obvious advantage of tracking both the inner and outer lip contours is that additional shape information is available for input to the recognition engine. However, a more substantial benefit is that demarcation of the inner lip contour facilitates reasoning about the visibility of articulators inside the mouth. In the simplest case, a crude estimate of the presence/absence of the teeth can be obtained by computing the average grey-level inside the mouth as shown in figure 7.14. Although, having identified the inner mouth region, it would be natural
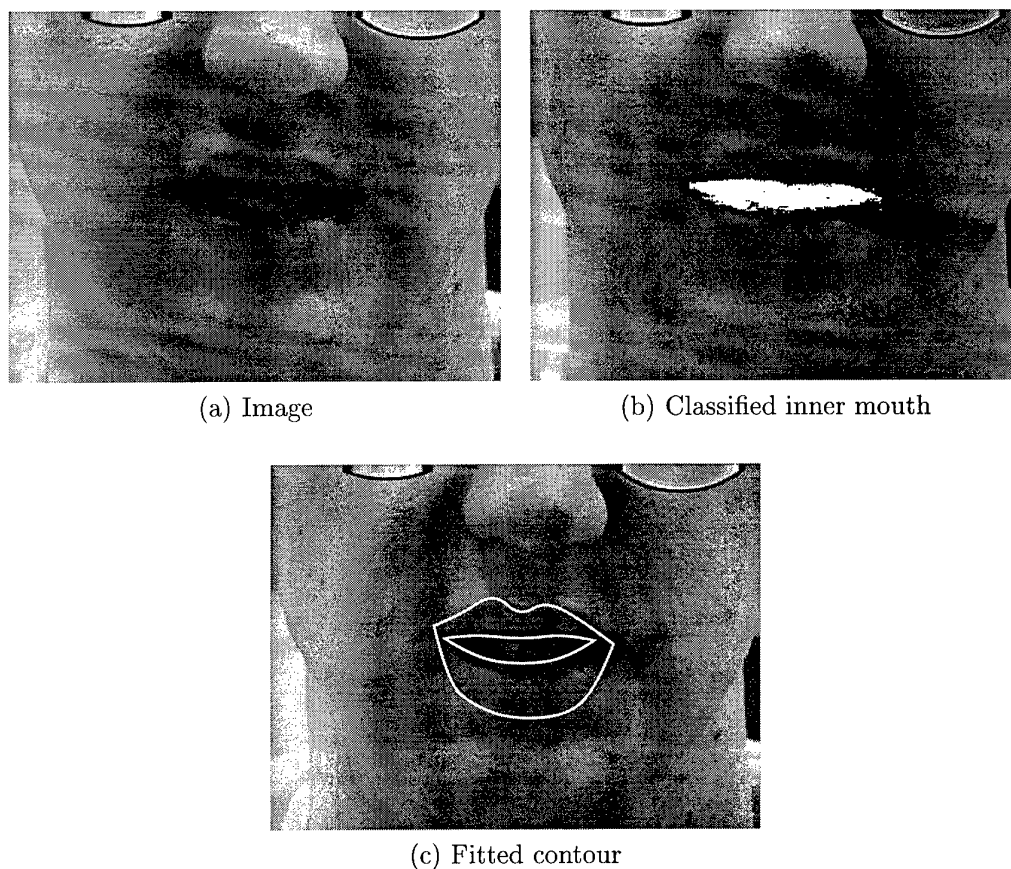
(a) Image

(b) Classified inner mouth

(c) Fitted contour

Figure 7.13: *Classification of the inner mouth and the resultant fitted contour. By modelling the distributions of the colour intensities inside the mouth and of the lips, the inner mouth region can be successfully segmented from its surround (b). Although identification of the inner mouth region for this image is particularly good, the tracking framework effectively handles misclassifications (see text).*

to employ more focused image processing, possibly utilising Gaussian mixture models, as discussed previously, in conjunction with a spatially-dependent pattern classifier, to make more detailed judgements about the position of the teeth and tongue. However, as shown in figure 7.15, even the average intensity inside the mouth provides additional information over shape information alone.

Tracking of both the inner and outer lip contours has been accomplished on more than four minutes of connected speech and in all cases tracking was stable — reliably following the lip outlines. However, the same level of accuracy attained for the outer lip contour over a long sequence has not yet been achieved using the inner-outer contour tracker. The principal difficulty is that, at times, the tracker is slow in responding to some of the more rapid lip
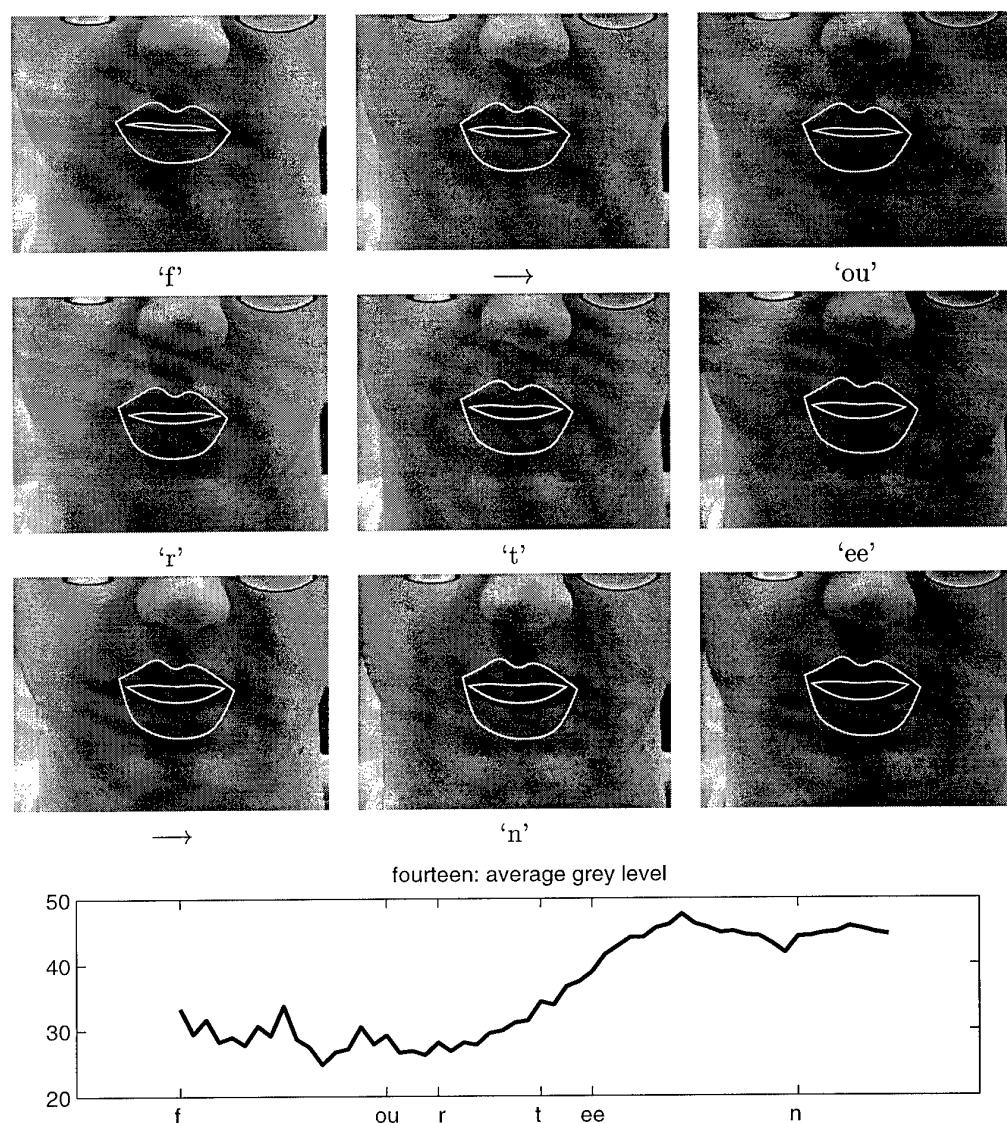
Figure 7.14: *Tracking the word "fourteen" on a colour sequence. Snapshots taken approximately every 120 ms. Both the inner and outer contour trackers follow their respective lips throughout the sequence. Minor tracking errors are visible at the mouth corners and along the upper right portion of the outer lip contour, but the tracker successfully handles both the nearly closed mouth in 'f' and the appearance of teeth in the 't'. The underlying plot shows the average grey-level intensity inside the mouth. Naturally, for audio-visual speech recognition, information more beneficial to speech recognition, such as the fact that only the lower teeth are visible and that the tongue is completely obscured, would need to be extracted; however, even the average intensity is informative as the onset of the 't' is clearly evident by the rapid increase in intensity.*
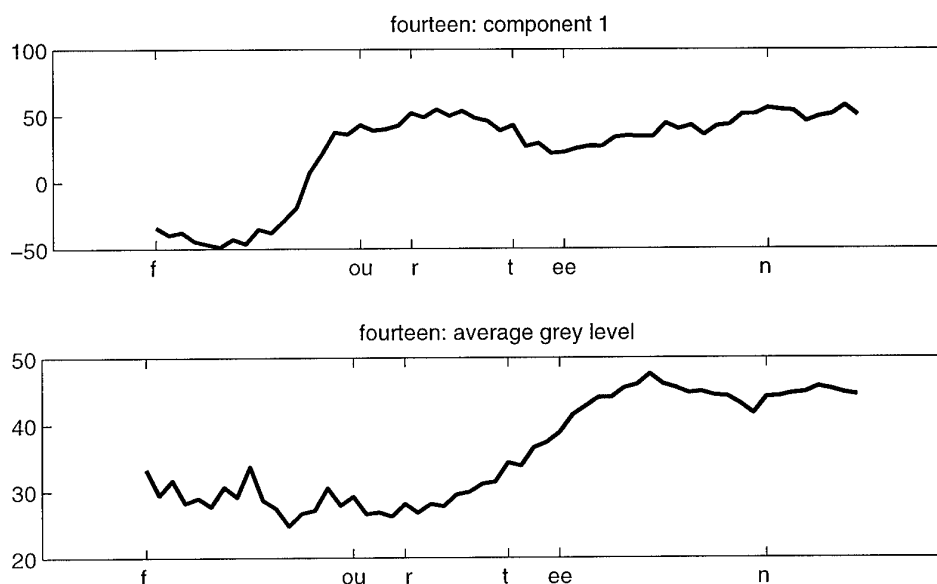
Figure 7.15: *The average grey-level intensity inside the mouth complements the shape information. Shape information corresponding to the degree of mouth opening (top) and inner mouth region information (bottom) for the tracked sequence of "fourteen" (figure 7.14) is shown. The opening of the mouth is clearly captured by the shape signal, while the average grey-level intensity identifies the onset of the 't'.*

movements. This is illustrated in figure 7.16 where the tracker lags slightly during the rapid 'e' to 'v' transition in "seventeen". The articulatory movements of the 'v' sound (frame 4 in the figure) consist of a rapid closing of the mouth in concert with an inward curling of the lower lip. During this movement, the outer contour begins to contract, although the inner contour becomes momentarily situated over the lower lip rather than the inner mouth. However, as evidenced by the subsequent frames, the tracker quickly recovers during the completion of the fricative 'v' and tracks accurately to the end of the word.

A plot of the mouth height (figure 7.17) illustrates that the tracker has adequately captured the principal articulatory movements of "seventeen". For instance, the initial opening of the mouth for 'e' and the rapid closure of 'v' are readily identifiable. One can also see the characteristically slow opening of the mouth in 'teen'. However, the tracking error in the 'e' to 'v' transition manifests itself in a plot of the "inner-mouth" intensity. The sharp increase in intensity around the 'v' is due to the inner mouth tracker being positioned over the lower lip rather than the inner mouth (recall frame 4 of figure 7.16). A more informative measure of the inner-mouth articulators using a tongue/teeth/lips/dark-region classifier could be used to cure the symptoms of the problem; however, this example
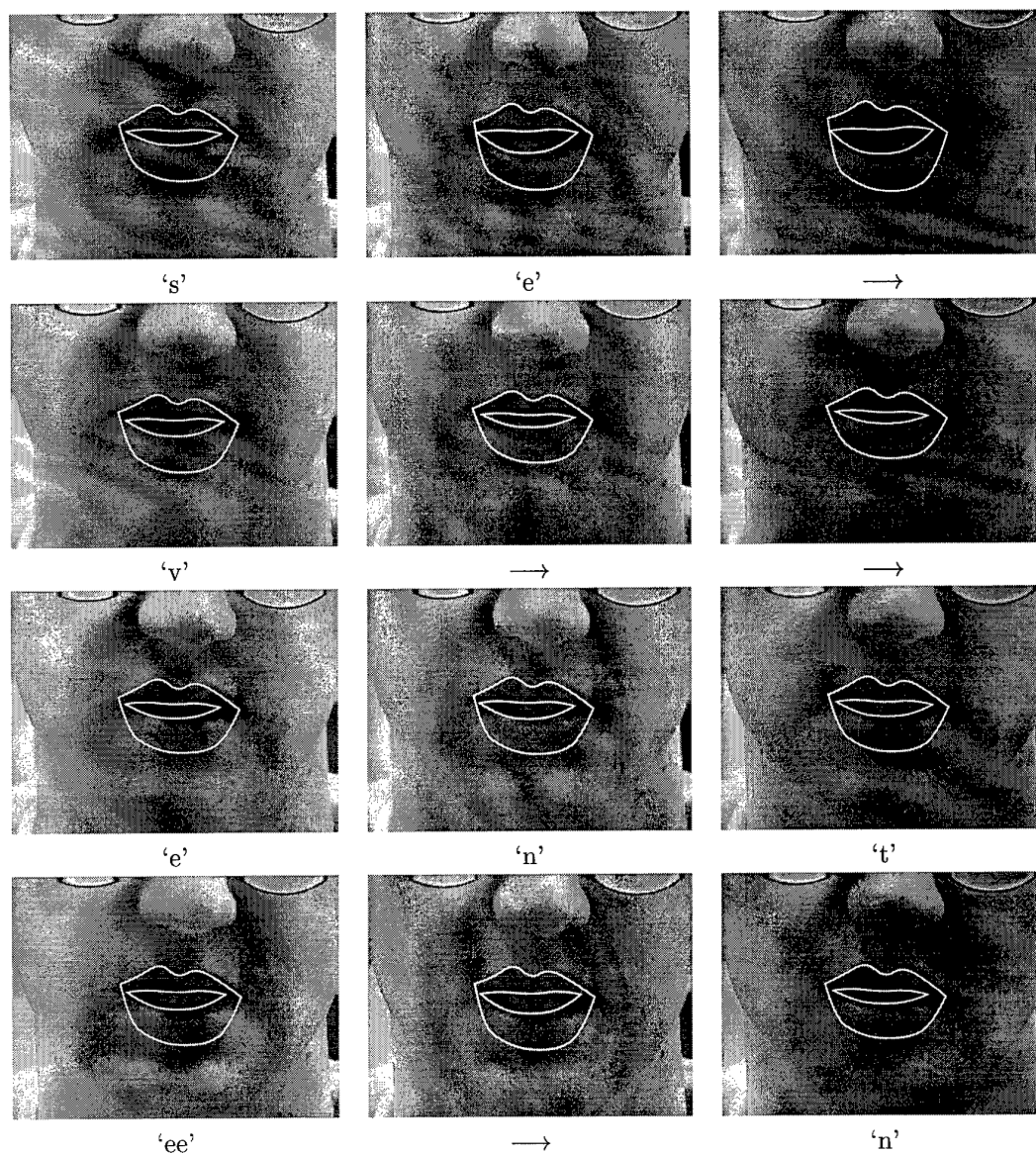
Figure 7.16: *Tracking the word "seventeen" on a colour sequence. Both the inner and outer trackers follow the general pattern of the deforming lips. However, the tracker lags slightly during the rapid 'e' to 'v' transition (frame 4) — the outer contour begins to contract, but the inner contour becomes momentarily situated over the lower lip rather than the inner mouth. The tracker quickly recovers in the next two frames during the completion of the fricative 'v' and tracks accurately to the end of the word.*

illustrates that when using intensity information the recognition engine must be able to generalise over illumination changes as well as compensate for tracking errors.

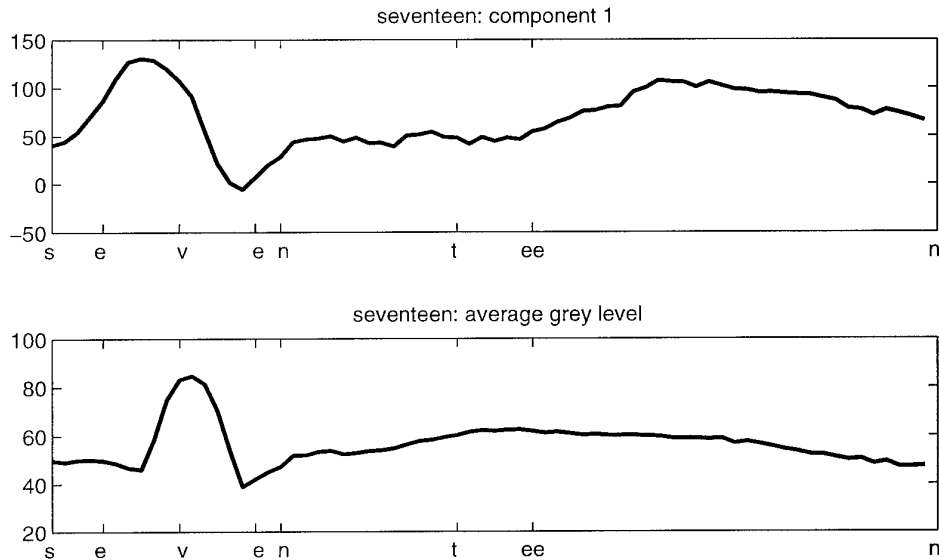**Figure 7.17**: *A momentary tracking error can result in the extraction of the intensity data of the lower lip rather than the inner mouth. Shape information corresponding to the degree of mouth opening (top) and inner mouth region information (bottom) for the tracked sequence of "seventeen" (figure 7.16) is shown. Although the overall motion of the lips has been captured (top), the momentary tracking error surrounding the 'v' (frames 4 and 5 in figure 7.16) results in misidentification of the inner mouth region and the extraction of lip intensity data which produces the hump in the bottom graph.*

Further work remains to be accomplished on this tracker in order to achieve the same degree of accurate, long term performance as was attained with the outer lip contour tracker. One area where improvements could be made concerns the shape and motion models employed. Presently a nine-dimensional shape-space learnt from principal components analysis is used. Potentially, expanding to a larger shape-space would enable tracking of some of the more subtle lip movements. Further, the sluggishness of the tracker in following some of the more rapid lip movements suggests that the dynamical models learnt were too rigid. Increasing the number of rapid movements in the training sequences might result in dynamics capable of following the quick, agile movements. There also is room for improvement in the measurement models used. Currently, image features are searched for only along normals to the contour. Potentially, measurement routines that make use of the information in the entire *region* surrounding the lips would lead to increased tracking accuracy.

7.7 Conclusions

The trackers presented in this chapter represent an advancement in state-of-the-art lip tracking. Previously, only Petajan et al. [106, 107] could successfully track unadorned lips within a real-time framework; however, their system relies on having a clear view of the nostrils, whereas the system presented here needs no such assistance. The approach taken here relies on the combination of several powerful techniques to permit accurate, real-time tracking. Comprehensive shape and motion models, which provide global structure and motion coherence, enable the use of focused image feature detection methods. These feature detectors, which employ Bayesian discriminant analysis techniques to colour images, provide for fast, accurate, identification of the boundary between the lips and their surround. The result is a robust, outer lip contour tracker. Further, the successful tracking of both the inner and outer lip contours presents a gateway to further exploration of the benefits of lipreading. Tracking of the inner lip contour provides supplementary shape information and enables detailed judgements to be made about the positioning of the teeth and tongue. These additional information sources permit more effective capture of the linguistic information inherent in visual speech and should lead to more effective audio-visual speech recognisers.

8

Conclusions and Future Work

This thesis has addressed several of the outstanding problems associated with the development of practical audio-visual speech recognition systems. A central aim of this work has been the providing of solutions to the difficult real-time lip tracking problem. A disciplined approach to this problem was taken. First, it was shown that the use of learnt shape and motion models enabled the tracking of rapidly moving, articulating lips when lipstick was used to enhance the contrast of the lips. It was then shown that the principal obstacle to accurate identification of the lips in unadorned situations is the poor contrast between the lips and facial skin. It was in these settings that the comprehensive modelling central to the dynamic contour tracking framework could be exploited to its fullest. Firstly, shape models were used to provide global structure to the lip contour and restrict its deformations to shape spaces characteristic of the speaker. Secondly, motion models were used which captured the temporal coherence of articulating lips. Lastly, statistical models of the grey-level appearance around the lips were employed which captured the information necessary for identifying the lip boundary. It was then shown that the fusion of these three modelling approaches resulted in accurate tracking of unadorned lips.

Having successfully tracked the lips, it was next essential to demonstrate that the tracked contours did indeed capture some of the linguistically informative aspects of the visual speech signal. Two audio-visual recognition systems, one which used a dynamic time warping pattern matching algorithm and the other which used continuous density Hidden Markov Models, were constructed for this purpose. The visual recognition features consisted

of shape parameters which were obtained from the outer lip contour. Experiments using the two recognisers were conducted on isolated-word vocabularies over a range of acoustic signal-to-noise ratios using additive Gaussian noise. The results of the experiments demonstrated that shape parameters obtained from accurately tracked lip contours could be used to provide robust speech recognition in the presence of high levels of interfering noise. Further, depending on the severity of the degradation of the acoustic channel, incorporation of the visual speech components resulted in error-rate reductions on the order of 44%. In other experiments, it was shown that the visual signal provided benefit even when the acoustic recogniser was trained and tested on speech at known noise levels. The results obtained establish dynamic contour tracking and computer lipreading as effective methods for improving the accuracy and robustness of automatic speech recognition systems.

Despite the generous improvement afforded by the visual shape parameters in adverse acoustic conditions, there was only a slight increase in recognition performance when the acoustic signal was clean. It was reasoned that additional visual information in the form of knowledge of the inner mouth region, to include the teeth and tongue, may be needed to increase performance in these environments. With this in mind, new trackers were developed which made use of the increased discriminating potential inherent in colour images of the face. A novel application of Fisher's Linear Discriminant Analysis was presented which enabled accurate identification of the lip-skin boundary and was shown to be robust to environmental variations. Further, since the learning of the Fisher discriminant was done off-line, the real-time performance of the outer lip contour tracker was not compromised. Accurate demarcation of the inner mouth contour was also attained despite considerable variations in the appearance of the mouth due to the varying presence of the teeth and tongue. Mixtures of multi-variate Gaussians enabled precise modelling of the colour intensities inside the mouth. The resultant inner-outer lip contour tracker permitted extraction of the region inside the mouth, thus enabling more detailed judgements to be made about the presence and position of the teeth and tongue. Although no recognition experiments were accomplished using these trackers, they provide a gateway to further exploration of the benefits attainable by augmenting acoustic speech recognisers with visual speech cues.

8.1 Future Work

Despite the successes achieved, the challenges provided by the audio-visual speech recognition problem afford many opportunities for further research.

8.1.1 Model Transfer

One area of research that has relevance to lip tracking, as well as many other motion capture problems, such as hand tracking, face and gesture recognition, and performance driven animation, concerns the shape and motion models inherent in the dynamic contour tracking framework. In this thesis, accurate lip tracking was demonstrated on several different speakers; however, the tracking was *speaker dependent*, that is, each lip tracker employed shape and motion models that were particular to the given speaker. Although a basic tenet of modelling is to incorporate as much prior knowledge as possible, from a commercial perspective it may be impractical to require that new models be developed from scratch every time a new user is to be tracked. This is particularly applicable in the lip tracking case where often the models are learnt from tracked sequences obtained using lipstick. It would be far preferable to devise a method where the shape and motion models could somehow be transferred from one speaker to another or derived from generic models.

One possible solution to this might be to develop a library of lip templates, shape spaces, and motion models from people with a wide range of lip shapes and visual articulatory patterns. During initialisation some method would be needed to determine the most appropriate models for the given speaker (in the simplest case, possibly the user himself could select the models). A more elegant approach might be to employ a variation of the *speaker adaptation* methods used by commercial speech recognisers. In speech recognition, speaker adaptation is accomplished by first learning speaker-independent models from training on a large number of speakers. Training utterances are then acquired from a new speaker by having him read from a pre-set list of sentences. Typically, the speaker-independent models are then customised by adapting the general models using a learnt mapping from the speaker-specific spectral data to that in the speaker-independent models [111]. A similar strategy might prove useful for customising generic lip-shape and motion models to an individual talker. The difficulty will be how best to incorporate the results of the tracking adaptation step, where the generic models are used to track the new user. This will likely be complicated by the fact that initially the lip tracking will be less than perfect and manual intervention may be required in order to provide ground-truth to the model adaption algorithm.

8.1.2 Region-based Measurement Routines

Another interesting area for potential research, particularly applicable to the tracking of both the inner and outer lip contours, concerns the use of improved *feature measurement*

models. Since the Kalman filter requires that both the dynamical process and the measurement densities be described analytically by Gaussian densities, the observation (measurement) densities for all of the feature detection methods presented in this thesis were necessarily modelled as Gaussians. In particular, mutually independent 1D observations were made along normals to the lip contours to obtain features corresponding to the measured lip position. Potentially, the employment of more powerful feature detection methods which exploit the spatial dependencies between the search lines or utilise *region* information surrounding the contour would lead to enhanced tracking performance.

A tracking framework capable of handling such measurement densities has recently been developed by Blake and Isard [13, 71]. Their algorithm, termed CONDENSATION (for Conditional Density Propagation, since it propagates the entire probability distribution of object position and shape over time, rather than just estimates for the mean and covariance, as is the case in Kalman-filter trackers), effectively handles observation densities of arbitrary form. One of the strengths of the CONDENSATION tracker is that it tracks objects by generating curve configurations and then scoring the hypothesised curve according to the amount of image support. As currently implemented [71], feature measurements are taken along 1D normals to the hypothesised curve as is done in the Kalman-filter trackers employed here; however, the CONDENSATION framework can be extended to handle arbitrary measurement densities including those using region information. For example, rather than using only edge information, the measurements could be obtained from the entire region bounded by the contours, for instance the upper lip or inner mouth. One such measure might consist of image "moments" for a given region R_i,

$$\mathbf{z}_i = \frac{1}{A(R_i)} \int_{R_i} g(x,y)\mathbf{I}(x,y)dA$$

where the measurements $\mathbf{z}_i$ are normalised by the area of the region $A(R_i)$. Here $g(x,y)$ is a (spatially varying) weighting function corresponding to the different moments, eg. $g(x,y) = 1$ gives the mean intensity of the region. Further, it may also be possible to exploit the spatial texture information in and around the lips using Markov Random Fields.

One can also imagine a situation where it would be desirable to use a combination of normal-based and region measurements. For example, in the case of tracking both the inner and outer lip contours, Fisher detectors could be used to identify the outer lip contour, while region measurements might be used to accurately identify the inner lip contour. Further, with this approach, judgements of the positioning of the teeth and tongue might be a natural by-product of tracking.

8.1.3 Shape and Mouth Region Recognition Features

Although it has been established that parameters representing the lip shape are a rich source of information for audio-visual speech recognition, further work is needed to extract additional information from the facial images and to assess the information content of these additional visual features. The dual inner-outer lip contour tracker of chapter 7 represents a first step in achieving this goal. As a first attempt, recognition experiments could be accomplished to assess the increase in recognition performance obtained via the incorporation of shape parameters from the tracked inner lip contour. Further experiments could focus on the added contribution provided by utilising colour intensity information from inside the mouth. It may also be possible to develop teeth and tongue detectors which can provide supplemental recognition information without being susceptible to variations in illumination, as is often the case when using the pixel data directly. The Gaussian mixture modelling used in the inner-outer lip contour tracker could be used to provide initial estimates of the teeth and tongue location, although it is likely that additional spatial modelling of the inner mouth region will be required to provide robust estimates.

8.1.4 High-level Knowledge Sources

This thesis has focused on the integration of the audio and visual information at the lowest level in the speech recognition process, that is, during feature extraction, although it is important not to ignore the role that cognitive skill or "intelligence" plays in natural-language understanding. As humans, often it is our intimate familiarity with the English language, its rules, grammar, and the content of the message, that permits us to "fill in" words or pieces of words that may have been obscured aurally and/or visually in a sentence [43, 98]. This capability, coupled with our innate ability to blend incoming sensory information, may account for our unmatched capacity to recognise spoken language. The research presented here does not incorporate any of these higher level knowledge sources; it is certainly possible, and even likely, that the full benefit provided by visual information will only be realised in systems utilising these higher level knowledge sources.

Bibliography

[1] A. Adjoudani and C. Benoit. On the integration of auditory and visual parameters in an HMM-based ASR. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 461–472. NATO Scientific Affairs Division, Sep 1995.

[2] K. J. Astrom and B. Wittenmark. *Computer Controlled Systems*. Addison Wesley, 1984.

[3] D.H. Ballard and C.M. Brown. *Computer Vision*. Prentice– Hall, New Jersey, 1982.

[4] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.

[5] R.H. Bartels, J.C. Beatty, and B.A. Barsky. *An Introduction to Splines for use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann, 1987.

[6] B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. In *Proc. 6th Int. Conf. on Computer Vision*, 1998. in press.

[7] B. Bascle and R. Deriche. Region tracking through image sequences. In *Proc. 5th Int. Conf. Computer Vision*, pp. 302–307, Boston, Jun 1995.

[8] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Stat.*, 41(1):164–171, 1970.

[9] C. Benoit, B. Guiard-Marigny, B. Le Goff, and A. Adjoudani. Which components of the face do humans and machines best speechread ? In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 315–328. NATO Scientific Affairs Division, Sep 1995.

[10] Michael Black and Yaser Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proc. 5th Int. Conf. Computer Vision*, pp. 374–381, 1995.

[11] A. Blake, R. Curwen, and A. Zisserman. Affine-invariant contour tracking with automatic control of spatiotemporal scale. In *Proc. 4th Int. Conf. Computer Vision*, pp. 66–75, 1993.

[12] A. Blake, R. Curwen, and A. Zisserman. A framework for spatio-temporal control in the tracking of visual contours. *Int. J. Computer Vision*, 11(2):127–145, 1993.

[13] A. Blake and M. Isard. Condensation — conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 1997. in press.

[14] A. Blake and M.A. Isard. 3D position, attitude and shape input using video tracking of hands and lips. In *Proc. Siggraph*, pp. 185–192. ACM, 1994.

[15] A. Blake, M.A. Isard, and D. Reynard. Learning to track the visual motion of contours. *Artificial Intelligence*, 78:101–134, 1995.

[16] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 557–560, Minneapolis, 1993.

[17] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 669–672, Adelaide, 1994.

[18] C. Bregler and S.M. Omohundro. Surface learning with applications to lipreading. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, volume 6. Morgan Kaufmann Publishers, 1994.

[19] C. Bregler and S.M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proc. 5th Int. Conf. Computer Vision*, pp. 494–499, Boston, Jun 1995.

[20] C. Bregler, S.M. Omohundro, J. Shi, and Y. Konig. Towards a robust speechreading dialog system. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 409–424. NATO Scientific Affairs Division, Sep 1995.

[21] J. S. Bridle, K. M. Ponting, M. D. Brown, and A. W. Borrett. A noise compensating spectrum distance measure applied to automatic speech recognition. In *Proceedings of the Institute of Acoustics*, volume 6, pp. 307–314, 1984.

[22] N.M. Brooke. Mouth shapes and speech. In V. Bruce and M. Burton, editors, *Processing Images of the Face*, pp. 20–40. Ablex publishing corporation, 1992.

[23] N.M. Brooke. Talking heads and speech recognisers that can see: The computer processing of visual speech signals. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 351–372. NATO Scientific Affairs Division, Sep 1995.

[24] N.M. Brooke and E.D. Petajan. Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. In *Proc. Intl. Conf. Speech Input and Output: Techniques and Applications*, pp. 104–109. Science Education and Technology Division of the IEE, Mar 1986.

[25] N.M. Brooke and Q. Summerfield. Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics*, 11:63–76, 1983.

[26] N.M Brooke, M. Tomlinson, and R. Moore. Automatic speech recognition that includes visual speech cues. *Proceedings of the Institute of Acoustics*, 16(5):15–22, Autumn 1994.

[27] R. Campbell. Lipreading. In A.W. Young and H.D. Ellis, editors, *Handbook of Research on Face Processing*, pp. 187–201. Elsevier Science Publishers, 1989.

[28] J.F. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[29] D. Chandramoha and P. L. Silsbee. A multiple deformable template approach for visual speech recognition. In *Proc. Fourth Intl. Conf. on Spoken Language Processing*, volume 1, pp. 42–45, 1996.

[30] R. Cipolla and A. Blake. The dynamic analysis of apparent contours. In *Proc. 3rd Int. Conf. Computer Vision*, pp. 616–625, 1990.

[31] Arnon Cohen. Karhunen-loeve expansions. In *Biomedical Signal Processing*, volume 2, pp. 66–75. CRC Press, 1986.

[32] R. Cole, L. Hirschmann, L. Atlas, et al. The challenge of spoken language systems: Research directions for the nineties. *IEEE Trans. on Speech and Audio Processing*, 3(1):1–20, 1995.

[33] T.F. Cootes and C.J. Taylor. Active shape models. In *Proc. British Machine Vision Conf.*, pp. 265–275, 1992.

[34] T.F. Cootes, C.J. Taylor, A. Lanitis, D.H. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. In *Proc. 4th Int. Conf. Computer Vision*, pp. 242–246, 1993.

[35] L. Couch. *Digital and Analog Communication Systems*. Macmillan, 1993.

[36] Jill D. Crisman. Color region tracking for vehicle guidance. In Andrew Blake and Alan Yuille, editors, *Active Vision*, chapter 7. The MIT Press, 1992.

[37] R. Curwen. *Dynamic and Adaptive Contours*. PhD thesis, University of Oxford, 1993.

[38] R. Curwen and A. Blake. Dynamic contours: real-time active splines. In A. Blake and A. Yuille, editors, *Active Vision*, pp. 39–58. MIT, 1992.

[39] B. Dalton, R. Kaucic, and A. Blake. Automatic speechreading using dynamic contours. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 373–382. NATO Scientific Affairs Division, Sep 1995.

[40] A. Dempster, M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, B(39):1–38, 1977.

[41] R. Deriche. Using Canny's criteria to derive a recursively implemented optimal edge detector. *Int. J. Computer Vision*, 1:167–187, 1987.

[42] B. Dodd. Lip-reading, phonological coding and deafness. In B. Dodd and R Campbell, editors, *Hearing By Eye: The Psychology of Lip Reading*, pp. 177–189. Erlbaum, 1987.

[43] B. Dodd and R. Campbell. *Hearing By Eye: The Psychology of Lip Reading*. Erlbaum, 1987.

[44] R. Dorf. *Modern Control Systems*. Addison-Wesley, 1992.

[45] P. Duchnowski, M. Hunke, D. Busching, U. Meier, and A. Waibel. Toward movement-invariant automatic lip-reading and speech recognition. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, 1995.

[46] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[47] E. K. Finn and A. A. Montgomery. Automatic optically based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, 1988.

[48] Kathleen E. Finn. *An Investigation of Visible Lip Information to be Used in Automated Speech Recognition.* PhD thesis, Georgetown University, 1986.

[49] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE. Trans. Computers,* C-22(1), 1973.

[50] M.M. Fleck, D.A. Forsyth, and C. Bregler. Finding naked people. In *Proc. 4th European Conf. Computer Vision,* pp. 593–602, Cambridge, England, Apr 1996.

[51] M.J.F. Gales and S. Young. An improved approach to the Hidden Markov Model decomposition of speech and noise. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing,* pp. 233–239, San Franciso, Mar 1992.

[52] M.J.F. Gales and S. Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication,* 12(3):231–239, 1993.

[53] M.J.F. Gales and S. Young. Robust continuous speech recognition using parallel model combination. *IEEE Trans. on Speech and Audio Processing,* 4(5):352–359, Sept 1996.

[54] A. Gelb, editor. *Applied Optimal Estimation.* MIT Press, Cambridge, MA, 1974.

[55] A.J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading.* PhD thesis, George Washington University, Sep 1993.

[56] A.J. Goldschen, O.N. Garcia, and E.D. Patajan. Rationale for phoneme-viseme mapping and feature selection in visual speech recognition. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications,* pp. 505–515. NATO Scientific Affairs Division, Sep 1995.

[57] M.S. Gray, Javier R. Movellan, and T.J Sejnowski. Dynamic features for visual speechreading: A systematic comparison. In Mozer, Jordan, and Petsche, editors, *Advances in Neural Information Processing Systems,* volume 9. 1996.

[58] M.S. Grewal and A.P. Andrews. *Kalman Filtering: Theory and Practice.* Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[59] T. Guiard-Marigny, A. Adjoudani, and C. Benoit. 3D models of the lips and jaw for visual speech synthesis. In J. van Santen, editor, *Progress in Speech Synthesis.* Springer-Verlag, 1996.

[60] C. Harris. Tracking with rigid models. In A. Blake and A. Yuille, editors, *Active Vision,* pp. 59–74. MIT, 1992.

[61] C.G. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pp. 147–151, 1988.

[62] M. Hennecke, D. Stork, and K. Prasad. Visionary speech: Looking ahead to practical speechreading systems. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 331–350. NATO Scientific Affairs Division, Sep 1995.

[63] Marcus E. Hennecke, K. Venkatesh Prasad, and David G. Stork. Using deformable templates to infer visual speech dynamics. In *28th Asilomar Conference on Signals, Systems, and Computers*, pp. 578–582. IEEE Computer Society Press, November 1994.

[64] Marcus E. Hennecke, K. Venkatesh Prasad, and David G. Stork. Automatic speech recognition system using visual signals. In *29th Asilomar Conference on Signals, Systems, and Computers*. IEEE Computer Society Press, 1995.

[65] H. Hermansky. Perceptual linear predictive (PLP) anaysis of speech. *J. Acoustical Society of America*, 87(4):1738–1752, Apr 1990.

[66] I.J. Hirsh, H. Davis, S.R. Silverman, E.G. Reynolds, E. Eldert, and R.W. Benson. Development of material for speech audiometry. *J. Speech and Hearing Disorders*, 17:321–337, 1952.

[67] J.N. Holmes. *Speech Synthesis and Recognition*. Van Nostrand Reinhold (UK) Co. Ltd., 1988.

[68] B.K.P. Horn. *Robot Vision*. McGraw-Hill, NY, 1986.

[69] B.K.P. Horn and B.G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[70] X.D. Huang, Y. Arika, and M.A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, 1990.

[71] M.A. Isard and A. Blake. Visual tracking by stochastic propagation of conditional density. In *Proc. 4th European Conf. Computer Vision*, pp. 343–356, Cambridge, England, Apr 1996.

[72] G. Jacob and A. Noble. Evaluating a robust contour tracker on echocardiographic sequences. In *Medical image understanding and analysis*, Jul 1997. in press.

[73] O.L.R. Jacobs. *Introduction to control theory.* Oxford University Press, 1993.

[74] B.H. Juang. Speech recognition in adverse environments. *Computer Speech and Language,* 5:275–294, 1991.

[75] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. 1st Int. Conf. Computer Vision,* pp. 259–268, 1987.

[76] R. Kaucic and A. Blake. Accurate, real-time, unadorned lip tracking. In *Proc. 6th Int. Conf. on Computer Vision,* 1998. in press.

[77] R. Kaucic, B. Dalton, and A. Blake. Real-time liptracking for audio-visual speech recognition applications. In *Proc. 4th European Conf. Computer Vision,* pp. 376–387, Cambridge, England, Apr 1996.

[78] R. Kaucic, D. Reynard, and A. Blake. Real-time liptrackers for use in audio-visual speech recognition. In *IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication,* pp. 31–36, London, England, Nov 1996.

[79] R. Kjeldsen and J. Kender. Finding skin in colour images. In *International Conference on Automatic Face and Gesture Recognition,* pp. 312–317, Oct 1996.

[80] D.H. Klatt. A digital filterbank for spectral masking. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing,* pp. 573–576, 1976.

[81] J.J. Koenderink and A.J. Van Doorn. Affine structure from motion. *J. Optical Soc. America A.,* 8(2):337–385, 1991.

[82] E. Kreysig. *Advanced Engineering Mathematics.* Wiley, 1988.

[83] A. Lanitis, C.J. Taylor, and T.F. Cootes. A unified approach to coding and interpreting face images. In *Proc. 5th Int. Conf. Computer Vision,* pp. 368–373, 1995.

[84] B. Le Goff, T. Guiard-Marigny, and C. Benoit. Read my lips ... and my jaw! how intelligible are the components of a speaker's face? In *Proceedings of the 4th Eurospeech Conference,* pp. 291–294, Madrid, Spain, Sept 1995.

[85] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Proceedings of SIGGRAPH,* pp. 55–62, 1995.

[86] J. Leuttin, N. Thacker, and S. Beet. Active shape models for visual speech feature extraction. In *Proceedings NATO ASI Conference on Speechreading by Man and*

Machine: Models, Systems and Applications, pp. 383–390. NATO Scientific Affairs Division, Sep 1995.

[87] J. Leuttin, N. Thacker, and S. Beet. Visual speech recogntion using active shape models and hidden markov models. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, pp. 817–820, 1996.

[88] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell System Tech. Journal*, 62(4):1035–1074, 1983.

[89] M.W. Mak and W.G. Allen. Lip-motion analysis for speech segmentation in noise. *Speech Communication*, 14(3):279–296, 1994.

[90] M. Markow, H. Grady. Rylander III, and A. Welch. Real-time algorithm for retinal tracking. *IEEE Transactions on Biomedical Engineering*, 40(12):1269–1281, Dec 1993.

[91] D. Marr and E. Hildreth. Theory of edge detection. *Proc. Roy. Soc. London. B.*, 207:187–217, 1980.

[92] K. Mase and A. Pentland. Automatic lip-reading by optical flow analysis. Media Lab Report 117, MIT, 1991.

[93] D.W. Massaro. Speech perception by ear and eye. In B. Dodd and R Campbell, editors, *Hearing By Eye: The Psychology of Lip Reading*, pp. 53–83. Erlbaum, 1987.

[94] M. McGrath, A.Q. Summerfield, and N.M. Brooke. Roles of lips and teeth in lip reading vowels. In *Proceedings of the Institute of Acoustics*, volume 6, pp. 401–408, Windermere, autumn 1984.

[95] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[96] W.S. Meisel. ARPA workshop reports speech-recognition state-of-the-art. *Speech Recognition Update*, 20:1–20, Feb 1995.

[97] S. Menet, P. Saint-Marc, and G. Medioni. B-snakes: Implementation and application to stereo. In *Proceedings DARPA*, pp. 720–726, 1990.

[98] K. Mogford. Lip-reading in the prelingually deaf. In B. Dodd and R Campbell, editors, *Hearing By Eye: The Psychology of Lip Reading*, pp. 191–211. Erlbaum, 1987.

[99] A.A. Montgomery and P.L. Jackson. Physical characteristics of the lips underlying vowel lipreading performance. *J. Acoustical Society of America*, 73(6):2134–2144, 1983.

[100] Y. Moses, D. Reynard, and A. Blake. Determining facial expressions in real-time. In *Proc. 5th Int. Conf. Computer Vision*, pp. 296–301, Boston, Jun 1995.

[101] Javier R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT press, Cambridge, 1995.

[102] Ramakant Nevatia. A color edge detector and its use in scene segmentation. *IEEE Transactions on Systems, Man and Cybernetics*, 7(11):820–826, November 1977.

[103] J.P. Openshaw and J.S. Mason. A review of robust techniques for the analysis of degraded speech. In *Proc. IEEE Region 10 Conf. on Comp., Control, and Power Engr.*, pp. 329–332, 1993.

[104] E.D. Petajan. Automatic lipreading to enhance speech recognition. In *Proc. IEEE Global Telecom. Conf.*, pp. 265–272, Nov 1984.

[105] E.D. Petajan, B.J. Bischofy, D.A. Bodoff, and N.M. Brooke. An improved automatic lipreading system to enhance speech recognition. In E. Soloway, D. Frye, and S.B. Sheppard, editors, *Proc. Human Factors in Computing Systems*, pp. 19–25. ACM, 1988.

[106] E.D. Petajan and H.P. Graf. Robust face feature analysis for automatic speechreading and character animation. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 425–436. NATO Scientific Affairs Division, Sep 1995.

[107] E.D. Petajan and H.P. Graf. Robust face feature analysis fo automatic speechreading and character animation. In *International Conference on Automatic Face and Gesture Recognition*, pp. 357–362, Oct 1996.

[108] J.M. Pickett. *The Sounds of Speech Communication: A primer of Acoustic Phonetics and Speech Perception*. University Park Press, Baltimore, 1980.

[109] Francisco Pla, F. Juste, F. Ferri, and M. Vicens. Colour segmentation based on a light reflection model to locate citrus fruits for robotic harvesting. *Computers and Electronics in Agriculture*, 9(1):53–70, August 1993.

[110] A.B. Poritz. Hidden markov models: A guided tour. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, pp. 7–13, 1988.

[111] L. Rabiner and J. Bing-Hwang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.

[112] L. Rabiner and B.H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, pp. 4–16, Jan 1986.

[113] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

[114] B. Rao. Data association methods for tracking systems. In A. Blake and A. Yuille, editors, *Active Vision*, pp. 91–105. MIT, 1992.

[115] R.R. Rao and R.M. Mersereau. Lip modelling for visual speech recognition. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, pp. 587–590, 1994.

[116] D. Reisberg, J. McLean, and A. Goldfield. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd and R Campbell, editors, *Hearing By Eye: The Psychology of Lip Reading*, pp. 97–113. Erlbaum, 1987.

[117] D. Reynard, A. Wildenberg, A. Blake, and J. Marchant. Learning dynamics of complex motions from image sequences. In *Proc. 4th European Conf. Computer Vision*, pp. 357–368, Cambridge, England, Apr 1996.

[118] J. Robert-Ribes, M. Piquemal, J. Schwartz, and P. Escudier. Exploiting sensor fusion architectures and stimuli complementarily in av speech recognition. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 193–210. NATO Scientific Affairs Division, Sep 1995.

[119] H. Rossotti. *Colour: Why the World isn't Grey*. Princeton University Press, Princeton, N.J., 1983.

[120] S. Rowe and A. Blake. Statistical background modelling for tracking with a virtual camera. In *Proc. British Machine Vision Conf.*, volume 2, pp. 423–432, 1995.

[121] S.M. Rowe. *Robust feature search for active tracking*. PhD thesis, University of Oxford, 1996.

[122] S.M. Rowe and A. Blake. Statistical feature modelling for active contours. In *Proc. 4th European Conf. Computer Vision*, pp. 560–569, Cambridge, England, Apr 1996.

[123] T.J. Sejnowski, B.P. Yuhas, M.H. Goldstein, and R.E. Jenkins. Combining visual and acoustic speech signals with a neural network improves intelligibility. In D.S. Touretzky, editor, *Advances in Neural Information Processing 2*, pp. 232–239. Morgan Kaufman, 1990.

[124] Peter L. Silsbee and Alan C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4(5):337–351, Sept 1996.

[125] P.L. Silsbee. *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*. PhD thesis, University of Texas at Austin, May 1993.

[126] P.L. Silsbee and A.C. Bovik. Automatic lipreading. *Biomedical Sciences Instrumentation*, 29(3):415–422, 1993.

[127] K. Sobottka and I. Pitas. Segmentation and tracking of faces in color images. In *International Conference on Automatic Face and Gesture Recognition*, pp. 236–241, Oct 1996.

[128] D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. In *Proceedings International Joint Conference on Neural Networks*, volume 2, pp. 289–295, 1992.

[129] W.H. Sumby and I. Pollack. Visual contributions to speech intelligibility in noise. *J. Acoustical Society of America*, 26:212–215, 1954.

[130] Q. Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell, editors, *Hearing By Eye: The Psychology of Lip Reading*, pp. 3–51. Erlbaum, 1987.

[131] Q. Summerfield, A. MacLeod, M. McGrath, and M. Brooke. Lips, teeth and the benefits of lipreading. In A.W. Young and H.D. Ellis, editors, *Handbook of Research on Face Processing*, pp. 223–233. Elsevier Science Publishers, 1989.

[132] D. Terzopoulos and K. Waters. Analysis of facial images using physical and anatomical models. In *Proc. 3rd Int. Conf. Computer Vision*, pp. 727–732, 1990.

[133] M.J. Tomlinson, M.J. Russell, and N.M. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, 1996.

[134] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.

[135] A.P. Varga and R.K. Moore. Hidden Markov Model decomposition of speech and noise. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, pp. 845–848, 1990.

[136] S. V. Vasegi, P. N. Conner, and B. P. Milner. Speech modelling using cepstral-time feature matrices in hidden markov models. In *Proceedings of the Institute of Electrical Engineering*, volume 140, pp. 317–320, Oct 1993.

[137] S. V. Vasegi and B. P. Milner. Noise compensation methods for hidden markov model speech recognition in adverse environments. *IEEE Trans. on Speech and Audio Processing*, 5(1):11–21, 1997.

[138] A.J. Viterbi. Error bounds for convolution codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Information Theory*, IT-13:260–269, Apr 1967.

[139] M. Vogt. Fast matching of a dynamic lip model to color video sequences under regular illumination conditions. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 399–407. NATO Scientific Affairs Division, Sep 1995.

[140] A.P. Wildenberg. *Learning and Initialisation for Visual Tracking*. PhD thesis, University of Oxford, 1997.

[141] S.J. Young. The HTK Hidden Markov Model Toolkit: Design and philosophy. Technical Report CUED/F-INFENG/TR.152, Cambridge University, Dec 1993.

[142] B.P. Yuhas, M.H. Goldstein, and T.J. Sejnowski. Integration of acoustic and visual speech signals using neural nets. *IEEE Commun. Mag.*, pp. 65–71, Nov 1989.

[143] B.P. Yuhas, M.H. Goldstein, T.J. Sejnowski, and R.E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, 78(10):1658–1668, 1990.

[144] A. Yuille. Feature extraction from faces using deformable templates. *Int. Journal of Computer Vision*, 8(2):99–111, 1992.

[145] A. Yuille and P. Hallinan. Deformable templates. In A. Blake and A. Yuille, editors, *Active Vision*, pp. 20–38. MIT, 1992.

[146] A.L. Yuille, D.S. Cohen, and P.W. Hallinan. Feature extraction from faces using deformable templates. *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 104–109, 1989.

[147] V. Zue, J. Glass, D. Goodine, L. Hirschman, H. Leung, M. Phillips, J. Polifroni, and S. Seneff. From speech recognition to spoken language understanding: The development of the MIT SUMMIT and VOYAGER systems. In R.P. Lippman, J.E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing 3*, pp. 255–261. Morgan Kaufman, 1991.